Springer Theses
Recognizing Outstanding Ph.D. Research

Yong Cheng

Joint Training for Neural Machine Translation

清華大學出版社
TSINGHUA UNIVERSITY PRESS

Springer

Springer Theses

Recognizing Outstanding Ph.D. Research

Aims and Scope

The series "Springer Theses" brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student's supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today's younger generation of scientists.

Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

More information about this series at http://www.springer.com/series/8790

Yong Cheng

Joint Training for Neural Machine Translation

Doctoral Thesis accepted by
Tsinghua University, Beijing, China

清華大学出版社
TSINGHUA UNIVERSITY PRESS

Springer

Author
Dr. Yong Cheng
Google
Beijing, China

Supervisor
Prof. Wei Xu
Institute for Interdisciplinary
Information Sciences
Tsinghua University
Beijing, China

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-981-32-9747-0 ISBN 978-981-32-9748-7 (eBook)
https://doi.org/10.1007/978-981-32-9748-7

Jointly published with Tsinghua University Press
The print edition is not for sale in China. Customers from China please order the print book from: Tsinghua University Press.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Supervisor's Foreword

Recent years have witnessed the rapid development of neural machine translation (NMT), which have achieved tremendous success in both academia and industry. Significantly different from traditional statistical machine translation (SMT), NMT models the translation process by a holistic neural network based on an encoder-decoder framework. Dispensing with the combination of multiple translation components in SMT, the end-to-end training makes NMT excel traditional approaches greatly on the performance and efficiency. In his book, Dr. Cheng proposes a series of approaches to use two directional models. Different from standard NMT that usually describes the translation process in the one-way model without considering the other direction, Cheng's approaches focus on the interaction and collaboration between two directional models, enriching the learning information of the one-way model and complement each other. The book directly addresses three important challenges in NMT, including imprecise attentional alignment, data scarcity in parallel corpora, and ignoring bidirectional dependencies. Cheng proposes four innovative techniques to address the three challenges above:

1. He presents an agreement-based joint training approach to encourage two directional models to agree on word alignment matrices.
2. He proposes a semi-supervised learning approach to integrating monolingual corpora.
3. He further extends this idea to the low-resource NMT by bridging the source-to-target translation with a pivot language.
4. He also captures the bidirectional dependencies with an end-to-end bidirectional model. A contrastive learning approach further enhances the interaction between two directional models.

Cheng rigorously evaluated all the above techniques using state-of-the-art machine translation benchmarks and showed their superior performance. In addition to these new techniques, Cheng has provided a comprehensive survey on existing approaches, and thus the book is also good for newcomers to this field.

Beijing, China
June 2019

Prof. Wei Xu

Preface

Machine translation has achieved great success in the past few decades. The emergence and development of neural machine translation (NMT) have pushed the performance and practicality of machine translation to new heights. While NMT has obtained state-of-the-art results as a new paradigm, it still suffers from many drawbacks introduced by the new framework and machine translation itself.

The standard NMT usually builds a translation model from source to target. The modeling and training procedures in NMT are independent without the interaction with other NMT models such as an inverse translation model. In this book, we propose approaches to jointly training two directional NMT models, including the following topics:

1. Improving attentional mechanism: The attentional mechanism has proved to be effective in capturing long dependencies in NMT. However, due to the intricate structural divergence between natural languages, unidirectional attention-based models might only capture partial aspects of attentional regularities. We propose agreement-based joint training to encourage the two complementary models to agree on word alignment matrices on the same training data.
2. Incorporating monolingual corpora: NMT systems heavily rely on parallel corpora for parameter estimation. Since parallel corpora are usually limited in quantity, quality, and coverage, especially for low-resource languages, it is appealing to exploit monolingual corpora to improve NMT. We propose a semi-supervised approach for training NMT models on the concatenation of labeled (parallel corpora) and unlabeled (monolingual corpora) data. The semi-supervised approach uses an autoencoder to reconstruct monolingual corpora, in which the source-to-target and target-to-source translation models serve as the encoder and decoder, respectively.
3. Improving pivot-based translation: NMT systems suffer from the data scarcity problem for resource-scarce language pairs. Although this problem can be alleviated by exploiting a pivot language to bridge the source and target languages, the source-to-pivot and pivot-to-target translation models are usually independently trained. In this work, we introduce a joint training algorithm for

pivot-based NMT. We are committed to connecting two models closely to enable them to interact with each other.

4. Integrating bidirectional dependencies: The standard NMT only captures unidirectional dependencies to model the translation procedure from source to target. Nevertheless, the inverse information is explicitly available to reinforce the confidence of the translation process. We propose an end-to-end bidirectional NMT model to connect the source-to-target and target-to-source translation models, which opens up the interaction of parameters between two directional models. A contrastive learning approach is also adopted to further enhance the information sharing.

This book not only introduces four interesting research works that propose a novel idea of combining multiple NMT directional models but also covers the basic techniques of NMT and some potential research directions. It can make novice researchers enter the NMT field quickly and broaden their view for the advanced development of NMT.

Beijing, China
June 2019

Dr. Yong Cheng

Parts of This Book Have Been Published in the Following Articles

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. Neural machine translation with pivot languages. In International Joint Conference on Artificial Intelligence (IJCAI), 2017. (Reproduced with Permission from IJCAI Organization)

Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Agreement-based joint training for bidirectional attention-based neural machine translation. In International Joint Conference on Artificial Intelligence (IJCAI), 2016. (Reproduced with Permission from IJCAI Organization)

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. In Association for Computational Linguistics (ACL), 2016. (Reproduced with Permission from ACL Organization)

Contents

Chapter 1
Neural Machine Translation

Abstract The emergence of neural machine translation (NMT) has revolutionized the filed of machine translation. In the first section, we introduce the fundamental of NMT models. Then we study the advantages of NMT over traditional statistical machine translations (SMT), some of its existing challenges and recent research efforts. Finally, we summarize our four works that address several existing problems in NMT. In the second section, we follow attentional neural machine translation to use mathematical formula to define the overall procedure for NMT.

1.1 Introduction

With the rapid development of the economic globalization, language translation has acted as a communication bridge to connect people from different countries. Because professional interpreters are scarce and expensive, machine translation software tools, such as Google Translator [50] and Baidu Translator, have become convenient and essential in our daily life.

Machine translation is automated translation from one language to another using computer software. Many approaches have been developed to solve this big challenge in the past few decades since the field of machine translation appeared in Warren Weaver's Memorandum on Translation in 1949 [14]. The three of the most popular approaches are rule-based machine translation, example-based machine translation [37] and statistical machine translation [3, 29]. Among them, statistical machine translation (SMT), based on a noisy channel model, is the most widely used approach which has an advantage of the more efficient use of massive data resources. The log-linear model [38] enables SMT to easily incorporate well-designed sub-models, which advances its promotion.

Recent years witness the tremendous development of artificial intelligence brought by the application of neural network. Benefited from the success of end-to-end training based on neural network, neural machine translation (NMT) has been proposed as a new paradigm for machine translation [1, 9, 28, 46]. Without explicitly modeling latent structures that are vital for conventional SMT [3, 7, 29], NMT builds on an encoder-decoder framework: the encoder transforms a source-language sentence

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_1

into continuous-space representations through a recurrent neural network (RNN), from which the decoder generates a target-language sentence using another RNN. Compared with the conventional SMT, NMT has a number of obvious advantages. First, without relying on much domain knowledge and non-trivial sub-models, such as language model and translation model which play important roles in conventional SMT, NMT treats translation procedure as a sequence-to-sequence generation using a holistic neural network. We do not need to be excessively concerned about complex latent models. Second, unlike the complex decoder designed for conventional SMT, the decoder of NMT can generate a good translation just using a simple beam search algorithm. Third, the generalizability of neural network endows NMT with the ability to produce unseen word combinations which do not occur in the training set. These advantages make NMT achieve the state-of-art results compared with the current best translation systems [25, 33, 50].

While intrinsic advantages of NMT render it quickly successful as an advanced machine translation framework, it also brings some drawbacks which prevent NMT from further developing. We just list some part of them here. First, since NMT models encode a sentence into a fixed-length vector, the understanding and translation of long sentences are problematic although it's alleviated by enhancing recurrent neural networks (RNNs) with gate technique [9, 46] and attentional mechanisms [1]. Recent works also try to capture long dependencies through improving the accuracy of attentional mechanisms [5, 11, 17, 31, 33–36, 48]. Second, the number of target words is usually restricted to tens of thousands of the most frequent words due to the training complexity. It directly leads to the poor translation results containing unknown words which are out of vocabulary. Some techniques, like importance sampling [25], positional unknown model [33], byte pair encoding compression [43], and hybrid word-character models [10, 32], have been proposed to effectively handle the unknown word problem. Purely character-based neural machine translation is also a novel alternative to directly model translation process at the character level, which greatly reduces the training and memory complexities [12, 30]. Third, deep learning is a data-hungry learning technique which depends on massive labeled data. However, labeled data that refers to parallel corpora in NMT, is usually limited in quality, quantity and coverage, especially for low-resource languages. Thus it is appealing to exploit monolingual corpora. Recently, semi-supervised learning has drawn much attention for deep learning [13, 26] which can promote the utilization of the amounts of unlabelled data. Many works have also attempted to incorporate monolingual corpora to improve NMT [6, 20, 21, 40, 42, 51].

There has been the growth spurt for research works about NMT in recent years since NMT was first proposed in 2014 [1, 46]. Ranzato et al. [41] and Shen et al. [44] have proposed to directly optimize NMT with the BLEU evaluation metric instead of maximum likelihood. Exploiting better memory mechanism, like external memory, is successfully applied to NMT [47, 49]. The novel encoder-decoder framework also creates opportunities for NMT with numerous interesting directions. Multiple language translation [15, 18, 52], multi-modal translation [16, 22], knowledge-based translation [45], and zero-resources translation [4, 18, 27] are some of the representatives.

Our works mainly focus on resolving some existing problems in NMT through joint training approaches. The standard NMT usually models a source-to-target translation process on parallel sentences without considering the information of other directional models, such as a target-to-source model. Because the modeling and training procedures in NMT are completely independent with other translation models, we argue that the interaction with other models conduces to the performance of a single NMT model. Due to the structural divergence between natural languages, we deem that NMT models with different translation directions can be complementary and collaborative to solve some issues induced by just training a single model.

In our book, we propose approaches to jointly training two directional models to resolve existing problems in NMT from four aspects.

1. The first work is to improve the attentional mechanism of NMT, which can guide the decoder to find relevant source parts. We propose an agreement-based joint training approach for bidirectional NMT to make the attention learned more accurate [5]. The bidirectional NMT includes the source-to-target and target-to-source models. We encourage these two directional models to agree on word alignment matrices produced by the attentional mechanisms.
2. In the second work, we propose a semi-supervised approach for training NMT models on the concatenation of labeled (parallel corpora) and unlabeled (monolingual corpora) data through autoencoders, in which source-to-target and target-to-source models serve as the encoder and decoder respectively [6]. The main idea is to append a reconstruction term, which aims to reconstruct monolingual corpora.
3. We are also concerned with zero-resource NMT. For a source-to-target translation task with little or even no training data, we can introduce a pivot language to bridge the source and target languages. Nevertheless, the source-to-pivot and pivot-to-target translation models are usually independently trained. In this work, we introduce a joint training algorithm for pivot-based NMT to enhance the interaction of source-to-pivot and pivot-to-target translation models [4]. We propose three methods to connect the two models and enable them to interact with each other during training.
4. The fourth work presents an end-to-end bidirectional NMT model to capture bidirectional dependencies. We propose a soft connection approach to connect the source-to-target and target-to-source models which opens up the interaction between the two NMT models. We also propose an improved training approach, contrastive learning, to contrastively maximize the conditional probability to enhance the information sharing.

We compare our approaches with the state-of-the-art SMT and NMT systems on different language pairs, such as Chinese-English, German-English, French-English, and Spanish-English. Experimental results show that all our approaches can obtain significant performances and also achieve their respective goals.

1.2 Neural Machine Translation

Given a source-language sentence $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_M$ containing M words and a target-language sentence $\mathbf{y} = \mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N$ containing N words, the end-to-end neural machine translation (NMT) directly models the translation probability as a single and holistic neural network:

$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta}) \tag{1.1}$$

where $\boldsymbol{\theta}$ is a set of model parameters and $\mathbf{y}_{<n} = \mathbf{y}_1, \ldots, \mathbf{y}_{n-1}$ is a partial translation. There are several variations of NMT [1, 9, 46]. Here we mainly follow the framework, attention-based NMT, proposed by Bahdanau et al. [1], as is shown in Fig. 1.1.

The encoder-decoder framework [1, 9, 28, 46, 50] usually includes two parts, encoder and decoder. The encoder encodes a source sentences to a sequence of continuous vector representations, which can be either a recurrent neural network (RNN) [1, 46, 50] or a convolutional neural network (CNN) [8, 19]. The RNN is a primary choice for the encoder in most NMT systems. The decoder almost adopts an RNN to decode the target sentence based on the representations produced by the encoder.

The source sentence $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_M$ is transformed into a sequence of hidden states $\mathbf{h} = \mathbf{h}_1, \ldots, \mathbf{h}_m, \ldots, \mathbf{h}_M$ by a RNN. The $\mathbf{h}$ is calculated one by one as follows:

$$\mathbf{h}_m = f(\mathbf{x}_m, \mathbf{h}_{m-1}; \boldsymbol{\theta}) \tag{1.2}$$

where $\mathbf{h}_m$ is the hidden state of the m-th source word and $f(\cdot)$ is a non-linear function, such as $tanh$.

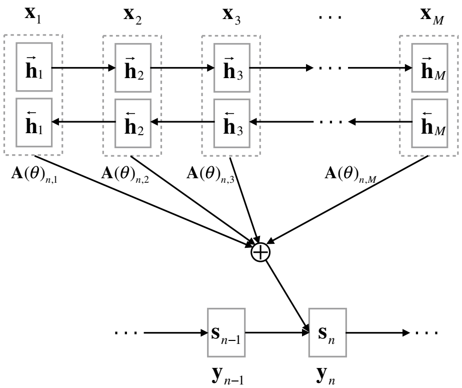

$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
...
$\mathbf{x}_M$
$\vec{\mathbf{h}}_1$
$\vec{\mathbf{h}}_2$
$\vec{\mathbf{h}}_3$
$\vec{\mathbf{h}}_M$
$\overleftarrow{\mathbf{h}}_1$
$\overleftarrow{\mathbf{h}}_2$
$\overleftarrow{\mathbf{h}}_3$
$\overleftarrow{\mathbf{h}}_M$
$\mathbf{A}(\theta)_{n,1}$
$\mathbf{A}(\theta)_{n,2}$
$\mathbf{A}(\theta)_{n,3}$
$\mathbf{A}(\theta)_{n,M}$
$\oplus$
$\mathbf{s}_{n-1}$
$\mathbf{s}_n$
$\mathbf{y}_{n-1}$
$\mathbf{y}_n$

Fig. 1.1 An illustration of attention-based NMT. The decoder generates a target hidden state $\mathbf{s}_n$ and its corresponding target word $\mathbf{y}_n$ given a source sentence $\mathbf{x}$. A bidirectional RNN is used to concatenate the forward and backward states as the hidden states of source words

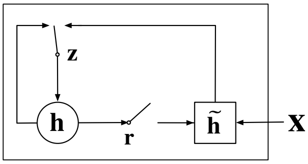


Fig. 1.2 An illustration of gated hidden activation function. It includes two gate mechanisms, in which the update gate $\mathbf{z}$ decides whether the hidden state is updated with the new hidden state $\tilde{\mathbf{h}}$ and the reset gate $\mathbf{r}$ controls whether the new hidden state $\tilde{\mathbf{h}}$ ignores the previous states. The input is $\mathbf{x}$ at the current time

As for the non-linear function $f(\cdot)$, there are mainly two types of extensions to the conventional activation function in RNN, which are the gated hidden unit proposed by Cho et al. [9] and the long short-term memory (LSTM) unit [23]. Both of them have the capacity for better learning and memorizing the long-term dependencies. It allows the gradients to flow backward easily for mitigating the effect of gradient vanishing [2, 24, 39].

The Fig. 1.2 shows how the gated hidden activation function is updated. The new hidden state $\mathbf{h}_m$ is computed through gated hidden units as follows:

$$\begin{aligned} \mathbf{h}_m &= f(\mathbf{x}_m, \mathbf{h}_{m-1}; \boldsymbol{\theta}) \\ &= (1 - \mathbf{z}_m) \circ \mathbf{h}_{m-1} + \mathbf{z}_m \circ \tilde{\mathbf{h}}_m \end{aligned} \tag{1.3}$$

where $\circ$ is an element-wise multiplication. $\mathbf{z}_m$ is the output of update gate and $\tilde{\mathbf{h}}_m$ is computed by:

$$\tilde{\mathbf{h}}_m = \tanh(W E_x[\mathbf{x}_m] + U[\mathbf{r}_m \circ \mathbf{h}_{m-1}] + b) \tag{1.4}$$

where $E_x[\mathbf{x}_m]$ is the embedding of the word $\mathbf{x}_m$, $\mathbf{r}_m$ is output of the reset gate, W,U are weighted matrices and b is a bias vector.

These gates of the gated hidden activation are the key advantages over the conventional RNN where reset gates $\mathbf{r}_m$ control how much and what information is left for the current time stamp from the previous time and update gates $\mathbf{z}_m$ make the current hidden unit maintain the information from previous time. They are computed by:

$$\mathbf{z}_m = \sigma(W_z E_x[\mathbf{x}_m] + U_z \mathbf{h}_{m-1} + b_z) \tag{1.5}$$

$$\mathbf{r}_m = \sigma(W_r E_x[\mathbf{x}_m] + U_r \mathbf{h}_{m-1} + b_r) \tag{1.6}$$

where σ is a logistic sigmoid function which ranges from 0 to 1. When $\mathbf{r}_m$ is close to zero, the hidden state $\tilde{\mathbf{h}}_m$ tends to ignore the previous hidden state $\mathbf{h}_{m-1}$ and is updated only with the current input $\mathbf{x}_m$. This unit makes the gated RNN capture short-term dependencies, which allows it drop any redundant information. When $\mathbf{z}_m$ approaches zero, the hidden state $\mathbf{h}_m$ focuses on memorizing the previous information $\mathbf{h}_{m-1}$ and ignore the current input $\tilde{\mathbf{h}}_m$. This unit enables the gated RNN to learn to capture long-term dependencies. These two gates collaborate to make the gated RNN adapt itself to remember the relevant information and ignore the redundant information. The LSTM network also acts similarly to this mechanism [23].

The conventional RNN reads the source sentence one by one from the beginning to the end, which leads to incapability to capture the following words. Hence, a bidirectional RNN is adopted to summarize not only the preceding words, but also the following words. It contains two unidirectional RNN, a forward RNN and a backward RNN. The forward RNN is used to read the source sentence in order starting from the first word and transform it into a sequence of the forward hidden state $\overrightarrow{\mathbf{h}}$. In contrast, the backward RNN encodes the source sentence starting from the last word to a sequence of the backward hidden state $\overleftarrow{\mathbf{h}}$. Thus, The encoder adopts the bidirectional RNN to obtain two directional sequences of hidden states given a source sentences. Then the forward hidden state $\overrightarrow{\mathbf{h}}$ is concatenated with the backward hidden state $\overleftarrow{\mathbf{h}}$ at its corresponding position for capturing the surrounding context information, that is.

$$\hat{\mathbf{h}}_m = [\overrightarrow{\mathbf{h}}_m; \overleftarrow{\mathbf{h}}_m] \tag{1.7}$$

where the dimension of $\hat{\mathbf{h}}_m$ is two times the size of the dimension of $\overrightarrow{\mathbf{h}}_m$ or $\overleftarrow{\mathbf{h}}_m$. The last hidden state in the backward is used to initialize the RNN of the decoder.

In the decoder, Bahdanau et al. [1] define the conditional probability in Eq. (1.1) as:

$$P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta}) \propto g(\mathbf{y}_{n-1}, \mathbf{s}_n, \mathbf{c}_n; \boldsymbol{\theta}) \tag{1.8}$$

where $g(\cdot)$ is a non-linear function, $\mathbf{s}_n$ is the hidden state corresponding to the n-th target word computed by

$$\mathbf{s}_n = f(\mathbf{s}_{n-1}, \mathbf{y}_{n-1}, \mathbf{c}_n; \boldsymbol{\theta}) \tag{1.9}$$

where $\mathbf{c}_n$ is a context vector for generating the n-th target word and f is the gated RNN as Eq. (1.2) indicated. We can make the decision for the generation of the current word $\mathbf{y}_n$ depending on the previous hidden state $\mathbf{s}_{n-1}$, the context vector $\mathbf{c}_n$, and the previous target word $\mathbf{y}_{n-1}$.

Compared with the encoder-decoder NMT proposed by Sutskever et al. [46], the $\mathbf{c}_n$ is a remarkable improvement which captures the relevant source information to assist generating the target word. Each annotation $\hat{\mathbf{h}}_i$ summarises the information of the whole source sentence with a strong focus on the local information surrounding the i-th source word. The context vector $\mathbf{c}_n$ is computed by:

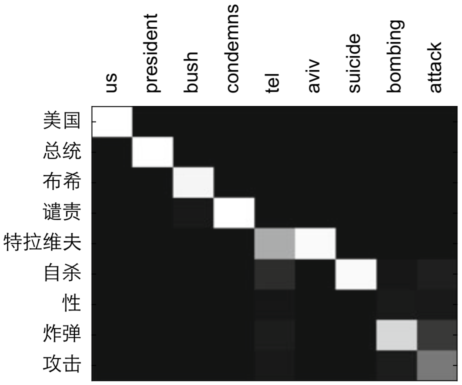


Fig. 1.3 An example alignment produced by NMT. Each pixel shows the weight $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ which measures the translation correlation between the m-th source word and the n-th target word, in grayscale (0: black, 1: white)

$$\mathbf{c}_n = \sum_{m=1}^{M} \mathbf{A}(\boldsymbol{\theta})_{n,m} \hat{\mathbf{h}}_m \tag{1.10}$$

We refer to $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{R}^{N \times M}$ as attentional mechanism, in which an element $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ reflects the contribution of the m-th source word $\mathbf{x}_m$ to generating the n-th target word $\mathbf{y}_n$ and transitioning to the next state $\mathbf{s}_n$. This decides the parts of the source sentences should be paid more attention to by the decoder. As Fig. 1.3 shows, it is clear that the attentional value $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ indicates the possibility of the translation for the n-th target word from the m-th source word. The $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ is calculated as:

$$\mathbf{A}(\boldsymbol{\theta})_{n,m} = \frac{\exp(a(\mathbf{s}_{n-1}, \hat{\mathbf{h}}_m; \boldsymbol{\theta}))}{\sum_{m'=1}^{M} \exp(a(\mathbf{s}_{n-1}, \hat{\mathbf{h}}_{m'}; \boldsymbol{\theta}))} \tag{1.11}$$

where $a(\mathbf{s}_{n-1}, \hat{\mathbf{h}}_m, \boldsymbol{\theta})$ measures how well $\mathbf{x}_m$ and $\mathbf{y}_n$ are aligned. The a denotes a multilayer network. Note that word alignment is treated as a function parameterized by $\boldsymbol{\theta}$ instead of a latent variable in attention-based NMT. It acts like the word alignment model in SMT. However, the alignment model in NMT is calculated in a soft way, which is in favour of the backpropagation of the gradients. So we can really achieve the end-to-end training for NMT. The attention mechanism relieves the burden of the encoder which has to memorize the entire source information in a fixed-length vector. Instead, it can dynamically and selectively retrieve useful information as additional context for the decoder.

Actually, the conditional probability of a target word $\mathbf{y}_n$ is defined as:

$$P(\mathbf{y}_n | \mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta}) = \frac{\exp^{(E_y[\mathbf{y}_n] W_o \mathbf{t}_n)}}{\sum_{y \in \mathcal{Y}} \exp^{(E_y[y] W_o \mathbf{t}_n)}} \tag{1.12}$$

where $\mathcal{Y}$ denotes the target vocabulary and $\mathbf{W}_o$ is a weighted matrix. $\mathbf{t}_n$ is computed by:

$$\mathbf{t}_n = \left[\max\left\{\tilde{\mathbf{t}}_{n,2j-1}, \tilde{\mathbf{t}}_{n,2j}\right\}\right]_{j=1,\ldots,l} \tag{1.13}$$

$$\tilde{\mathbf{t}}_n = U_o\mathbf{s}_n + V_o E_y[\mathbf{y}_{n-1}] + C_o\mathbf{c}_n \tag{1.14}$$

where U_o, V_o, and C_o are weighted matrices and $\tilde{\mathbf{t}}_n$ is a vector. We can find $\mathbf{t}_n$ is obtained by a single hidden maxout layer with the inputs, $\mathbf{s}_n$, $E_y[\mathbf{y}_{n-1}]$ and $\mathbf{c}_n$.

Given a set of training examples $\{\langle\mathbf{x}^{(s)}, \mathbf{y}^{(s)}\rangle\}_{s=1}^{S}$, the training algorithm aims to find the model parameters that maximize the likelihood of the training data:

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\{ \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \boldsymbol{\theta}) \right\} \tag{1.15}$$

References

1. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
2. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*.
3. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.
4. Cheng, Y., Liu, Y., Yang, Q., Sun, M., & Xu, W. (2016). Neural machine translation with pivot languages. arXiv:1611.04928.
5. Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Agreement-based joint training for bidirectional attention-based neural machine translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
6. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Semi-supervised learning for neural machine translation. In *Association for Computational Linguistics (ACL)*.
7. Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Association for Computational Linguistics (ACL)*.
8. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. arXiv:1409.1259.
9. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
10. Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Association for Computational Linguistics (ACL)*.
11. Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., & Haffari, G. (2016). Incorporating structural alignment biases into an attentional neural translation model. In *North American Association for Computational Linguistics (NAACL)*.
12. Costa-Jussa, M. R., & Fonollosa, J. A. R. (2016). Character-based neural machine translation. arXiv:1603.00810.
13. Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems (NIPS)*.

14. Delavenay, É. (1959). *La machine à traduire*, vol. 834. Presses universitaires de France.
15. Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. In *Association for Computational Linguistics (ACL)*.
16. Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., & Cohn, T. (2016). An attentional model for speech translation without transcription. In *North American Association for Computational Linguistics (NAACL)*.
17. Feng, S., Liu, S., Li, M., & Zhou, M. (2016). Implicit distortion and fertility models for attention-based encoder-decoder nmt model.
18. Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman Vural, F. T., & Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
19. Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2016). A convolutional encoder model for neural machine translation. arXiv:1611.02344.
20. Gulccehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. arXiv:1503.03535 [cs.CL].
21. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., & Ma, W.-Y. (2016). Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS)*.
22. Hitschler, J., Schamoni, S., & Riezler, S. (2016). Multimodal pivots for image caption translation. In *Association for Computational Linguistics (ACL)*.
23. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*.
24. Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Master's thesis, Institut fur Informatik, Technische Universitat, Munchen*.
25. Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Association for Computational Linguistics (ACL)*.
26. Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using lstm for region embeddings. arXiv:1602.02373.
27. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. arXiv:1611.04558.
28. Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Empirical Methods in Natural Language Processing (EMNLP)*.
29. Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *North American Association for Computational Linguistics (NAACL)*.
30. Ling, W., Trancoso, I., Dyer, C., & Black, A. W. (2015). Character-based neural machine translation. arXiv:1511.04586.
31. Liu, L., Utiyama, M., Finch, A., & Sumita, E. (2016). Neural machine translation with supervised attention.
32. Luong, M.-T., & Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. In *Association for Computational Linguistics (ACL)*.
33. Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
34. Meng, F., Lu, Z., Li, H., & Liu, Q. (2016). Interactive attention for neural machine translation.
35. Mi, H., Sankaran, B., Wang, Z., & Ittycheriah, A. (2016). Coverage embedding models for neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
36. Mi, H., Wang, Z., & Ittycheriah, A. (2016). Supervised attentions for neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
37. Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*.
38. Och, F. (2003). Minimum error rate training in statistical machine translation. In *Association for Computational Linguistics (ACL)*.
39. Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training recurrent neural networks. arXiv:1211.5063.

40. Ramachandran, P., Liu, P. J., & Le, Q. V. (2016). Unsupervised pretraining for sequence to sequence learning. arXiv:1611.02683.
41. Ranzato, M. A., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence level training with recurrent neural networks.
42. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving nerual machine translation models with monolingual data. In *Association for Computational Linguistics (ACL)*.
43. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Association for Computational Linguistics (ACL)*.
44. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Minimum risk training for neural machine translation. In *Association for Computational Linguistics (ACL)*.
45. Shi, C., Liu, S., Ren, S., Feng, S., Li, M., Zhou, M., Sun, X., & Wang, H. (2016). Knowledge-based semantic embedding for machine translation. In *Association for Computational Linguistics (ACL)*.
46. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.
47. Tang, Y., Meng, F., Lu, Z., Li, H., & Yu, P. L. H. (2016). Neural machine translation with external phrase memory. arXiv:1606.01792.
48. Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. In *Association for Computational Linguistics (ACL)*.
49. Wang, M., Lu, Z., Li, H., & Liu, Q. (2016). Memory-enhanced decoder for neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
50. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., & Macherey, K. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.
51. Zhang, J., & Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
52. Zoph, B., & Knight, K. (2016). Multi-source neural translation. In *North American Association for Computational Linguistics (NAACL)*.

Chapter 2
Agreement-Based Joint Training for Bidirectional Attention-Based Neural Machine Translation

Abstract The attentional mechanism has proven to be effective in improving end-to-end neural machine translation. However, due to the intricate structural divergence between natural languages, unidirectional attention-based models might only capture partial aspects of attentional regularities. In this chapter, we propose agreement-based joint training for bidirectional attention-based end-to-end neural machine translation. Instead of training source-to-target and target-to-source translation models independently, our approach encourages the two complementary models to agree on word alignment matrices on the same training data. Experiments on ChineseEnglish and English-French translation tasks show that agreement-based joint training significantly improves both alignment and translation quality over independent training.

2.1 Introduction

End-to-end neural machine translation (NMT) is a newly proposed paradigm for machine translation [1, 4, 7, 18]. Without explicitly modeling latent structures that are vital for conventional statistical machine translation (SMT) [2, 3, 9], NMT builds on an *encoder-decoder* framework: the encoder transforms a source-language sentence into a continuous-space representation, from which the decoder generates a target-language sentence.

While early NMT models encode a source sentence as a fixed-length vector, [1] advocate the use of *attention* in NMT. They indicate that only parts of the source sentence have an effect on the target word being generated. In addition, the relevant parts often vary with different target words. Such an attentional mechanism has proven to be an effective technique in text generation tasks such as machine translation [1, 16] and image caption generation [19].

However, due to the structural divergence between natural languages, modeling the correspondence between words in two languages still remains a major challenge for NMT, especially for distantly-related languages. For example, Luong et al. [16] report that attention-based NMT lags behind the Berkeley aligner [12] in terms of alignment error rate (AER) on the English-German data. One possible reason is that

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_2

unidirectional attention-based NMT can only capture partial aspects of attentional regularities due to the non-isomorphism of natural languages.

In this work, we propose to introduce agreement-based learning [12, 13] into attention-based neural machine translation. The basic idea is to encourage source-to-target and target-to-source translation models to agree on word alignment on the same training data. This can be done by defining a new training objective that combines likelihoods in two directions as well as an agreement term that measures the consensus between word alignment matrices in two directions. Experiments on Chinese-English and English-French datasets show that our approach is capable of better accounting for attentional regularities and significantly improves alignment and translation quality over independent training.

2.2 Agreement-Based Joint Training

In attention-based NMT, the attentional mechanism enables end-to-end NMT to capture longer dependency, which vastly improves the translation performance for long sentences [1]. We refer to $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{R}^{N \times M}$ as *alignment matrix*, in which an element $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ reflects the contribution of the m-th source word $\mathbf{x}_m$ to generating the n-th target word $\mathbf{y}_n$[1]:

$$\mathbf{A}(\boldsymbol{\theta})_{n,m} = \frac{\exp(a(\mathbf{s}_{n-1}, \mathbf{h}_m, \boldsymbol{\theta}))}{\sum_{m'=1}^{M} \exp(a(\mathbf{s}_{n-1}, \mathbf{h}_{m'}, \boldsymbol{\theta}))} \tag{2.1}$$

where $a(\mathbf{s}_{n-1}, \mathbf{h}_m, \boldsymbol{\theta})$ measures how well $\mathbf{x}_m$ and $\mathbf{y}_n$ are aligned. Note that word alignment is treated as a function parameterized by $\boldsymbol{\theta}$ instead of a latent variable in attention-based NMT.

Word alignment is a vital resource for conventional statistical machine translation. In NMT, attention mechanism is similar to the word alignment model in SMT, and it also plays a key role in guiding to find the source part to be translated. For attention-based NMT, we can get a sequence of attention vector $\mathbf{A}(\boldsymbol{\theta})_{n,1}$, ..., $\mathbf{A}(\boldsymbol{\theta})_{n,M}$ when generating n-th target word. Given a parallel sentence pair, when we "force" decode attention-based NMT to translate a source sentence exactly to its target sentence, the best alignment matrix is obtained as Fig. 2.1 shows. In the traditional word alignment model [2], some words of source sentences usually act as "garbage collectors" that align too many target words. Similar to the issue existing in traditional word alignment model, the alignment model of attention-based NMT is not accurate as we expect, because some source words are usually aligned too many times and some words are not aligned at all. The appearance on attention-based NMT about this problem is that the sum value $\sum_{t=1}^{M} \alpha_{ti}$ at source position i is too large. In Fig. 2.1, the source word

[1] We denote the alignment matrix as $\mathbf{A}(\boldsymbol{\theta})$ instead of α in [1] to emphasize that it is a function parameterized by $\boldsymbol{\theta}$ and differentiable. Although $\mathbf{s}_n$ and $\mathbf{c}_n$ also depend on $\boldsymbol{\theta}$, we omit the dependencies for simplicity.

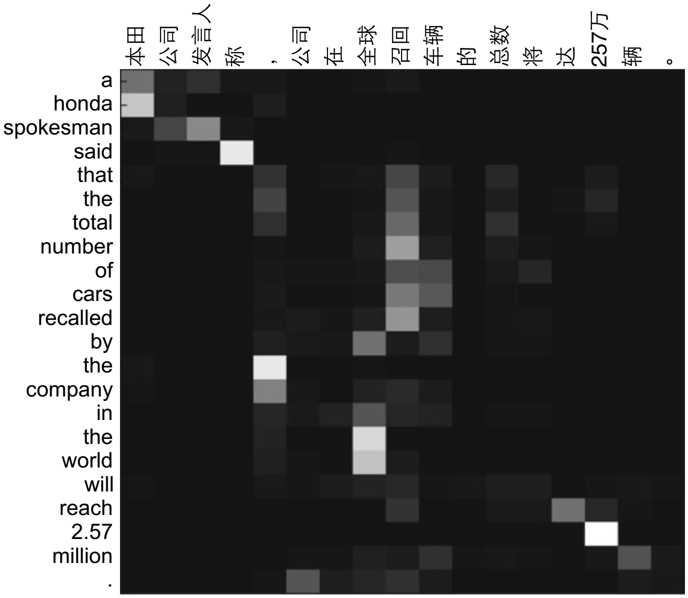


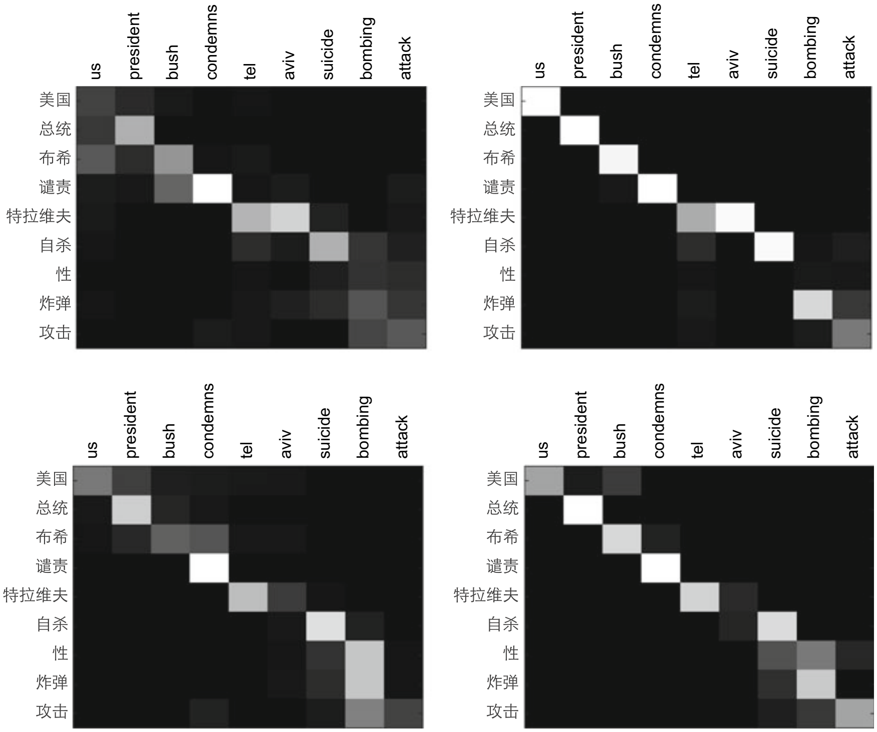
Fig. 2.1 A sample *alignment matrix* $\mathbf{A}(\boldsymbol{\theta})$ for Chinese-English sentence pair for attention-based NMT. The x-axis is source sentence (Chinese) and the y-axis is target sentence (English). The image denotes the *alignment matrix* $\mathbf{A}(\boldsymbol{\theta})$ where each pixel shows the weight $\mathbf{A}(\boldsymbol{\theta})_{n,m}$ in grayscale (0:black, 1:white)

"zhaohui" absorbs too much attention value from target side. This results out some words are not translated because of no attention and some words are translated many times. Luong et al. [15] verify that a better attentional mechanism contributes to the better performance of attention-based NMT.

Although the introduction of attention has advanced the state-of-the-art of NMT, it is still challenging for attention-based NMT to capture the intricate structural divergence between natural languages. We give one example to present our key idea. Figure 2.2a shows the Chinese-to-English (upper) and English-to-Chinese (bottom) alignment matrices for the same sentence pair. Both the two independently trained models fail to correctly capture the gold-standard correspondence: while the Chinese-to-English alignment assigns wrong probabilities to "us" and "bush", the English-to-Chinese alignment makes incorrect predictions on "condemns" and "bombing".

Fortunately, although each model only captures partial aspects of the mapping between words in natural languages, the two models seem to be complementary: the Chinese-to-English alignment does well on "condemns" and the English-to-Chinese alignment assigns correct probabilities to "us" and "bush". Therefore, combining the two models can hopefully improve alignment and translation quality in both directions.

In this work, we propose to introduce agreement-based learning [12, 13] into attention-based neural machine translation. The central idea is to encourage the



(a) independent training

(b) joint training

Fig. 2.2 Example alignments of **a** independent training and **b** joint training on a Chinese-English sentence pair. The first row shows Chinese-to-English alignments and the second row shows English-to-Chinese alignments. We find that the two unidirectional models are complementary and encouraging agreement leads to improved alignment accuracy

source-to-target and target-to-source models to agree on alignment matrices on the same training data. As shown in Fig. 2.2b, agreement-based joint training is capable of removing unlikely attention and resulting in more concentrated and accurate alignment matrices in both directions.

More formally, we train both the source-to-target attention-based neural translation model $P(\mathbf{y}|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}})$ and the target-to-source model $P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$ on a set of training examples $\{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$, where $\overrightarrow{\boldsymbol{\theta}}$ and $\overleftarrow{\boldsymbol{\theta}}$ are model parameters in two directions, respectively. The new training objective is given by

$$
\begin{aligned}
J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) = & \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)} | \mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}) \\
& + \sum_{s=1}^{S} \log P(\mathbf{x}^{(s)} | \mathbf{y}^{(s)}; \overleftarrow{\boldsymbol{\theta}}) \\
& - \lambda \sum_{s=1}^{S} \Delta\Big(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}, \overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\Big)
\end{aligned}
\tag{2.2}
$$

where $\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})$ is the source-to-target alignment matrix for the s-th sentence pair, $\overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})$ is the target-to-source alignment matrix for the same sentence pair, $\Delta(\cdot)$ is a loss function that measures the disagreement between two matrices, and λ is a hyper-parameter that balances the preference between likelihood and agreement.

For simplicity, we omit the dependency on the sentence pair and simply write the loss function as $\Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big)$. While there are many alternatives for quantifying disagreement, we use the following three types of loss functions in our experiments:

1. *Square of addition* (SOA): the square of the element-wise addition of corresponding matrix cells

$$
\begin{aligned}
& \Delta_{\text{SOA}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) \\
& = -\sum_{n=1}^{N} \sum_{m=1}^{M} \Big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} + \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n}\Big)^2
\end{aligned}
\tag{2.3}
$$

   Intuitively, this loss function encourages to increase the sum of the alignment probabilities in two corresponding matrix cells.

2. *Square of subtraction* (SOS): the square of the element-wise subtraction of corresponding matrix cells

$$
\begin{aligned}
& \Delta_{\text{SOS}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big) \\
& = \sum_{n=1}^{N} \sum_{m=1}^{M} \Big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} - \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n}\Big)^2
\end{aligned}
\tag{2.4}
$$

   Derived from the symmetry constraint proposed by Ganchev et al. [5], this loss function encourages that an aligned pair of words share close or even equal alignment probabilities in both directions.

3. *Multiplication* (MUL): the element-wise multiplication of corresponding matrix cells

$$\Delta_{\mathrm{MUL}}\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big)$$
$$= -\log \sum_{n=1}^{N} \sum_{m=1}^{M} \overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}})_{n,m} \times \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})_{m,n} \tag{2.5}$$

This loss function is inspired by the agreement term [12] and model invertibility regularization [11].

The decision rules for the two directions are given by

$$\overrightarrow{\boldsymbol{\theta}}^{*} = \operatorname*{argmax}_{\overrightarrow{\boldsymbol{\theta}}}\bigg\{ \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}) - \lambda \sum_{s=1}^{S} \Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big)\bigg\} \tag{2.6}$$

$$\overleftarrow{\boldsymbol{\theta}}^{*} = \operatorname*{argmax}_{\overleftarrow{\boldsymbol{\theta}}}\bigg\{ \sum_{s=1}^{S} \log P(\mathbf{x}^{(s)}|\mathbf{y}^{(s)}; \overleftarrow{\boldsymbol{\theta}}) - \lambda \sum_{s=1}^{S} \Delta\big(\overrightarrow{\mathbf{A}}^{(s)}(\overrightarrow{\boldsymbol{\theta}}), \overleftarrow{\mathbf{A}}^{(s)}(\overleftarrow{\boldsymbol{\theta}})\big)\bigg\} \tag{2.7}$$

Note that all the loss functions are differentiable with respect to model parameters. It is easy to extend the original training algorithm for attention-based NMT [1] to implement agreement-based joint training since the two translation models in two directions share the same training data.

2.3 Experiments

2.3.1 Setup

We evaluated our approach on Chinese-English and English-French machine translation tasks.

For Chinese-English, the training corpus from LDC consists of 2.56M sentence pairs with 67.53M Chinese words and 74.81M English words. We used the NIST 2006 dataset as the validation set for hyper-parameter optimization and model selection. The NIST 2002, 2003, 2004, 2005, and 2008 datasets were used as test sets. In the NIST Chinese-English datasets, each Chinese sentence has four reference English translations. To build English-Chinese validation and test sets, we simply "reverse" the Chinese-English datasets: the first English sentence in the four references as the source sentence and the Chinese sentence as the single reference translation.

For English-French, the training corpus from WMT 2014 consists of 12.07M sentence pairs with 303.88M English words and 348.24M French words. The concatenation of news-test-2012 and news-test-2013 was used as the validation set and news-test-2014 as the test set. Each English sentence has a single reference French translation. The French-English evaluation sets can be easily obtained by reversing the English-French datasets.

We compared our approach with two state-of-the-art SMT and NMT systems:

1. MOSES [8]: a phrase-based SMT system;
2. RNNSEARCH [1]: an attention-based NMT system.

For MOSES, we used the parallel corpus to train the phrase-based translation model and the target-side part of the parallel corpus to train a 4-gram language model using the SRILM [17]. We used the default system setting for both training and decoding.

For RNNSEARCH, we used the parallel corpus to train the attention-based NMT models. The vocabulary size is set to 30 K for all languages. We follow Jean et al. [6] to address the unknown word problem based on alignment matrices. Given an alignment matrix, it is possible to calculate the position of the source word to which is most likely to be aligned for each target word. After a source sentence is translated, each unknown word is translated from its corresponding source word. While Jean et al. [6] use a bilingual dictionary generated by an off-the-shelf word aligner to translate unknown words, we use unigram phrases instead.

In our system, we simply extends RNNSEARCH by replacing independent training with agreement-based joint training. The encoder-decoder framework and the attentional mechanism remain unchanged. The hyper-parameter λ that balances the preference between likelihood and agreement is set to 1.0 for Chinese-English and 2.0 for English-French. The training time of joint training is about 1.2 times longer than that of independent training for two directional models. We used the same unknown word post-processing technique as RNNSEARCH for our system.

2.3.2 Comparison of Loss Functions

We first compared the three loss functions as described in Sect. 2.3 on the validation set for Chinese-to-English translation. The evaluation metric is case-insensitive BLEU.

As shown in Table 2.1, the square of addition loss function (i.e., Δ_{SOA}) achieves the lowest BLEU among the three loss functions. This can be possibly attributed to the fact that a larger sum does not necessarily lead to increased agreement. For example, while $0.9 + 0.1$ hardly agree, $0.2 + 0.2$ perfectly does. Therefore, Δ_{SOA} seems to be an inaccurate measure of agreement.

The square of subtraction loss function (i.e, Δ_{SOS}) is capable of addressing the above problem by encouraging the training algorithm to minimize the difference between two probabilities: $(0.2 - 0.2)^2 = 0$. However, the loss function fails to distinguish between $(0.9 - 0.9)^2$ and $(0.2 - 0.2)^2$. Apparently, the former should be

Table 2.1 Comparison of loss functions in terms of case-insensitive BLEU scores on the validation set for Chinese-to-English translation

Loss	BLEU
Δ_{SOA}: square of addition	31.26
Δ_{SOS}: square of subtraction	31.65
Δ_{MUL}: multiplication	32.65

preferred because both models have high confidence in the matrix cell. It is unfavorable for two models agree on a matrix cell but both have very low confidence. Therefore, Δ_{SOS} is perfect for measuring agreement but ignores confidence.

As the multiplication loss function (i.e., Δ_{MUL}) is able to take both agreement and confidence into account (e.g., $0.9 \times 0.9 > 0.2 \times 0.2$), it achieves significant improvements over Δ_{SOA} and Δ_{SOS}. As a result, we use Δ_{MUL} in the following experiments.

2.3.3 Results on Chinese-English Translation

Table 2.2 shows the results on the Chinese-to-English (C → E) and English-to-Chinese (E → C) translation tasks.[2] We find that RNNSEARCH generally outperforms MOSES except for the C → E direction on the NIST08 test set, which confirms the effectiveness of attention-based NMT on distantly-related language pairs such as Chinese and English.

Agreement-based joint training further systematically improves the translation quality in both directions over independently training except for the E → C direction on the NIST04 test set.

2.3.4 Results on Chinese-English Alignment

Table 2.3 shows the results on the Chinese-English word alignment task. We used the TSINGHUAALIGNER evaluation dataset [14] in which both the validation and test sets contain 450 manually-aligned Chinese-English sentence pairs. We follow Luong et al. [16] to "force-decode" our jointly trained models to produce translations that match the references. Then, we extract only one-to-one alignments by selecting the source word with the highest alignment weight for each target word.

We find that agreement-based joint training significantly reduces alignment errors for both directions as compared with independent training. This suggests that intro-

[2] The scores for E → C is much lower than C → E because BLEU is calculated at the word level rather than character level.

Table 2.2 Results on the Chinese-English translation task. MOSES is a phrase-based statistical machine translation system. RNNSEARCH is an attention-based neural machine translation system. We introduce agreement-based joint training for bidirectional attention-based NMT. NIST06 is the validation set and NIST02-05, 08 are test sets. The BLEU scores are case-insensitive. "*": significantly better than MOSES ($p < 0.05$); "**": significantly better than MOSES ($p < 0.01$); "+": significantly better than RNNSEARCH with independent training ($p < 0.05$); "++": significantly better than RNNSEARCH with independent training ($p < 0.01$). We use the statistical significance test with paired bootstrap resampling [10]

System	Training	Direction	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08
MOSES	Indep.	C→E	32.48	32.69	32.39	33.62	30.23	25.17
		E→C	14.27	18.28	15.36	13.96	14.11	10.84
RNNSEARCH	Indep.	C→E	30.74	35.16	33.75	34.63	31.74	23.63
		E→C	15.71	20.76	16.56	16.85	15.14	12.70
	Joint	C→E	32.65^{++}	35.68^{**+}	34.79^{**++}	35.72^{**++}	32.98^{**++}	25.62^{*++}
		E→C	16.25^{*++}	21.70^{**++}	17.45^{**++}	16.98^{**}	15.70^{**+}	13.80^{**++}

Table 2.3 Results on the Chinese-English word alignment task. The evaluation metric is alignment error rate. "**": significantly better than RNNSEARCH with independent training ($p < 0.01$)

Training	C → E	E → C
Indep.	54.64	52.49
Joint	47.49^{**}	46.70^{**}

ducing agreement does enable NMT to capture attention more accurately and thus lead to better translations. Figure 2.2b shows example alignment matrices resulted from agreement-based joint training.

However, the error rates in Table 2.3 are still higher than conventional aligners that can achieve an AER around 30 on the same dataset. There is still room for improvement in attention accuracy.

2.3.5 Analysis of Alignment Matrices

We observe that a target word is prone to connect to too many source words in the alignment matrices produced by independent training. For example, in the lower alignment matrix of Fig. 2.2a, the third Chinese word "buxi" is aligned to three English words: "president", "bush", and "condemns". In addition, all the three alignment probabilities are relatively low. Similarly, four English words contribute to generating the last Chinese word "gongji": "condemns", "suicide", "boming", and "attack".

In contrast, agreement-based joint training leads to more concentrated alignment distributions. For example, in the lower alignment matrix of Fig. 2.2b, the third

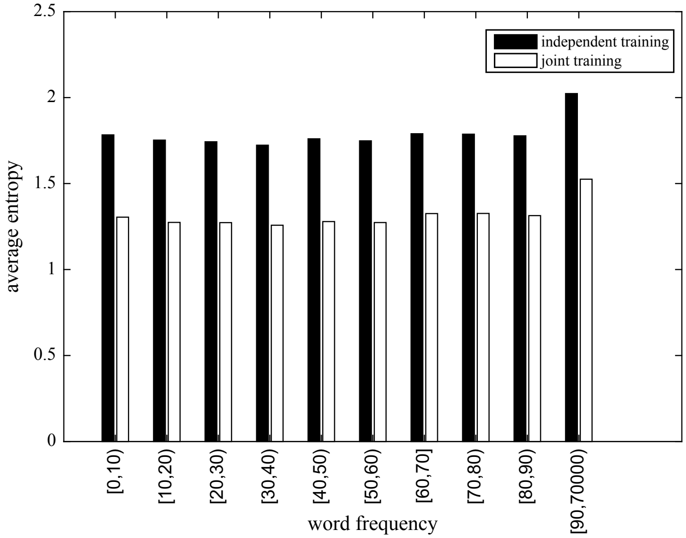


Fig. 2.3 Statistical results of independent and joint training in terms of average attention entropy

Chinese word "buxi" is most likely to be aligned to "bush". Likewise, the attention to the last Chinese word "gongji" now mainly focuses on "attack".

To measure the degree of concentration of attention, we define the *attention entropy* of a target word in a sentence pair as follows:

$$\mathrm{H}_{\mathbf{y}_n} = -\sum_{m=1}^{M} \mathbf{A}(\boldsymbol{\theta})_{n,m} \log \mathbf{A}(\boldsymbol{\theta})_{n,m} \tag{2.8}$$

Given a parallel corpus $D = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$, the *average attention entropy* is defined as

$$\tilde{\mathrm{H}}_y = \frac{1}{c(y, D)} \sum_{s=1}^{S} \sum_{n=1}^{N} \delta(\mathbf{y}_n^{(s)}, y) \mathrm{H}_{\mathbf{y}_n^{(s)}} \tag{2.9}$$

where $c(y, D)$ is the occurrence of a target word y on the training corpus D:

$$c(y, D) = \sum_{s=1}^{S} \sum_{n=1}^{N} \delta(\mathbf{y}_n^{(s)}, y) \tag{2.10}$$

As Fig. 2.3 shows, we can find that the entropy of the joint training is generally lower than that of the independent training. Table 2.4 gives the average attention entropy of example words on the Chinese-to-English translation task. We find that

Table 2.4 Comparison of independent and joint training in terms of average attention entropy (see Eq. 2.9) on Chinese-to-English translation

Word	Type	Freq.	Indep.	Joint
To	Preposition	High	2.21	1.80
And	Conjunction	High	2.21	1.60
The	Definite article	High	1.96	1.56
Yesterday	Noun	Medium	2.04	1.55
Actively	Adverb	Medium	1.90	1.32
Festival	Noun	Medium	1.55	0.85
Inspects	Verb	Low	0.29	0.02
Rebellious	Adjective	Low	0.29	0.02
Noticing	Verb	Low	0.19	0.01

the entropy generally goes downs with the decrease of word frequencies, which suggests that frequent target words tend to gain attention from multiple source words. Apparently, joint training leads to more concentrated attention than independent training. The gap seems to increase with the decrease of word frequencies.

2.3.6 Results on English-to-French Translation

Table 2.5 gives the results on the English-French translation task. While RNNSEARCH with independent training achieves translation performance on par with MOSES, agreement-based joint learning leads to significant improvements over both baselines. This suggests that our approach is general and can be applied to more language pairs.

Table 2.5 Results on the English-French translation task. The BLEU scores are case-insensitive. "**": significantly better than MOSES ($p < 0.01$); "++": significantly better than RNNSEARCH with independent training ($p < 0.01$)

System	Training	Direction	Dev.	Test
MOSES	Indep.	E→F	28.38	32.31
		F→E	28.52	30.93
RNNSEARCH	Indep.	E→F	29.06	32.69
		F→E	28.32	29.99
	Joint	E→F	29.86**++	33.45**++
		F→E	29.01**++	31.51**++

2.4 Summary

We have presented agreement-based joint training for bidirectional attention-based neural machine translation. By encouraging bidirectional models to agree on parametrized alignment matrices, joint learning achieves significant improvements in terms of alignment and translation quality over independent training. In the future, we plan to further validate the effectiveness of our approach on more language pairs.

References

1. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
2. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.
3. Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Association for Computational Linguistics (ACL)*.
4. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
5. Ganchev, K., Graça, J., Gillenwater, J., & Taskar, B. (2010). Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*.
6. Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Association for Computational Linguistics (ACL)*.
7. Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Empirical Methods in Natural Language Processing (EMNLP)*.
8. Koehn, P., & Hoang, H. (2007). Factored translation models. In *Empirical Methods in Natural Language Processing (EMNLP)*.
9. Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase-based translation. In *North American Association for Computational Linguistics (NAACL)*.
10. Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
11. Levinboim, T., Vaswani, A., & Chiang, D. (2015). Model invertibility regularization: Sequence alignment with or without parallel data. In *North American Association for Computational Linguistics (NAACL)*.
12. Liang, P., Taskar, B., Klein, D. (2006) Alignment by agreement. In *North American Association for Computational Linguistics (NAACL)*.
13. Liang, P., Klein, D., & Jordan, M. I. (2007). Agreement-based learning. In *Advances in Neural Information Processing Systems (NIPS)*.
14. Liu, Y., Sun, M. (2015). Contrastive unsupervised word alignment with non-local features. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
15. Luong, M.-T., Pham, H., Manning,C. D. (2015). Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
16. Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Association for Computational Linguistics (ACL)*.
17. Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*.

18. Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.
19. Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*.

Chapter 3
Semi-supervised Learning for Neural Machine Translation

Abstract While end-to-end neural machine translation (NMT) has made remarkable progress recently, NMT systems only rely on parallel corpora for parameter estimation. Since parallel corpora are usually limited in quantity, quality, and coverage, especially for low-resource languages, it is appealing to exploit monolingual corpora to improve NMT. We propose a semi-supervised approach for training NMT models on the concatenation of labeled (parallel corpora) and unlabeled (monolingual corpora) data. The central idea is to reconstruct the monolingual corpora using an autoencoder, in which the sourceto-target and target-to-source translation models serve as the encoder and decoder, respectively. Our approach can not only exploit the monolingual corpora of the target language, but also of the source language. Experiments on the ChineseEnglish dataset show that our approach achieves significant improvements over state-of-the-art SMT and NMT systems.

3.1 Introduction

End-to-end neural machine translation (NMT), which leverages a single, large neural network to directly transform a source-language sentence into a target-language sentence, has attracted increasing attention in recent several years [2, 10, 20]. Free of latent structure design and feature engineering that are critical in conventional statistical machine translation (SMT) [4, 5, 13], NMT has proven to excel in modeling long-distance dependencies by enhancing recurrent neural networks (RNNs) with the gating [6, 8, 20] and attention mechanisms [2].

However, most existing NMT approaches suffer from a major drawback: they heavily rely on parallel corpora for training translation models. This is because NMT directly models the probability of a target-language sentence given a source-language sentence and does not have a separate language model like SMT [2, 10, 20]. Unfortunately, parallel corpora are usually only available for a handful of research-rich languages and restricted to limited domains such as government documents and news reports. In contrast, SMT is capable of exploiting abundant target-side monolingual corpora to boost fluency of translations. Therefore, the unavailability of

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_3

large-scale, high-quality, and wide-coverage parallel corpora hinders the applicability of NMT.

As a result, several authors have tried to use abundant monolingual corpora to improve NMT. Gulccehre et al. [7] propose two methods, which are referred to as shallow fusion and deep fusion, to integrate a language model into NMT. The basic idea is to use the language model to score the candidate words proposed by the translation model at each time step or concatenating the hidden states of the language model and the decoder. Although their approach leads to significant improvements, one possible downside is that the network architecture has to be modified to integrate the language model.

Alternatively, Sennrich et al. [17] propose two approaches to exploiting monolingual corpora that is transparent to network architectures. The first approach pairs monolingual sentences with dummy input. Then the parameters of encoder and attention model are fixed when training on these pseudo parallel sentence pairs. In the second approach, they first train a neural translation model on the parallel corpus and then use the learned model to translate a monolingual corpus. The monolingual corpus and its translations constitute an additional pseudo parallel corpus. Similar ideas have also been suggested in conventional SMT [3, 21]. Sennrich et al. [17] report that their approach significantly improves translation quality across a variety of language pairs.

In this paper, we propose semi-supervised learning for neural machine translation. Given labeled (i.e., parallel corpora) and unlabeled (i.e., monolingual corpora) data, our approach jointly trains source-to-target and target-to-source translation models. The self-training is a straightforward semi-supervised learning scheme where the model is bootstrapped with additional data labelled by a credible prediction model. However, the learning process can be potentially biased if the initial model is degenerate: wrong model predictions are reinforced over time [11]. The key idea of our method is to append a reconstruction term to the training objective, which aims to reconstruct the observed monolingual corpora using an autoencoder. In the autoencoder, the source-to-target and target-to-source models serve as the encoder and decoder, respectively. As the inference is intractable, we propose to sample the full search space to improve the efficiency. Specifically, our approach has the following advantages:

1. *Transparent to network architectures*: our approach does not depend on specific architectures and can be easily applied to arbitrary end-to-end NMT systems.
2. *Both the source and target monolingual corpora can be used*: our approach can benefit NMT not only using target monolingual corpora in a conventional way, but also the monolingual corpora of the source language.

Experiments on Chinese-English NIST datasets show that our approach results in significant improvements in both directions over state-of-the-art SMT and NMT systems.

3.2 Semi-supervised Learning for Neural Machine Translation

3.2.1 *Supervised Learning*

Given a parallel corpus $\mathcal{D} = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$, the standard training objective in NMT is to maximize the likelihood of the training data:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}) \tag{3.1}$$

where $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$ is a neural translation model and $\boldsymbol{\theta}$ is a set of model parameters. $\mathcal{D}$ can be seen as *labeled* data for the task of predicting a target sentence $\mathbf{y}$ given a source sentence $\mathbf{x}$.

As $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$ is modeled by a single, large neural network, there does not exist a separate target language model $P(\mathbf{y}; \boldsymbol{\theta})$ in NMT. Therefore, parallel corpora have been the only resource for parameter estimation in most existing NMT systems. Unfortunately, even for a handful of resource-rich languages, the available domains are unbalanced and restricted to government documents and news reports. Therefore, the availability of large-scale, high-quality, and wide-coverage parallel corpora becomes a major obstacle for NMT.

3.2.2 *Autoencoders on Monolingual Corpora*

It is appealing to explore the more readily available, abundant monolingual corpora to improve NMT. Let us first consider an *unsupervised* setting: how to train NMT models on a monolingual corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$?

Our idea is to leverage *autoencoders* [18, 22]: (1) *encoding* an observed target sentence into a latent source sentence using a target-to-source translation model and (2) *decoding* the source sentence to reconstruct the observed target sentence using a source-to-target model. For example, as shown in Fig. 3.1b, given an observed English sentence “Bush held a talk with Sharon”, a target-to-source translation model (i.e., encoder) transforms it into a Chinese translation “bushi yu shalong juxing le huitan” that is unobserved on the training data (highlighted in grey). Then, a source-to-target translation model (i.e., decoder) reconstructs the observed English sentence from the Chinese translation.

More formally, let $P(\mathbf{y}|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}})$ and $P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$ be *source-to-target* and *target-to-source* translation models respectively, where $\overrightarrow{\boldsymbol{\theta}}$ and $\overleftarrow{\boldsymbol{\theta}}$ are corresponding model parameters. An autoencoder aims to reconstruct the observed target sentence via a latent source sentence:

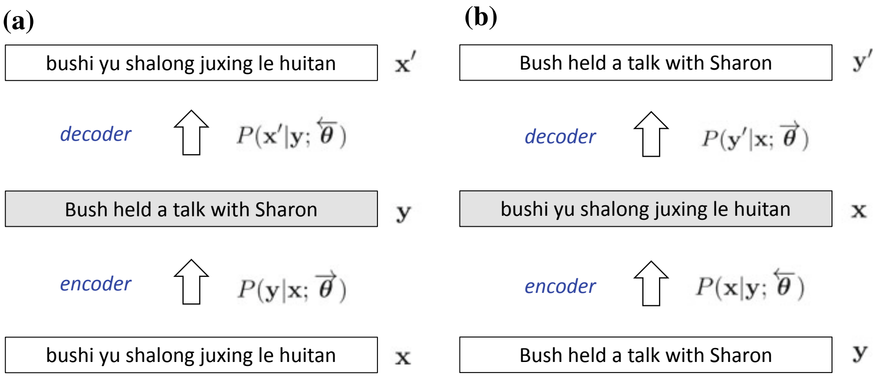

(a)
bushi yu shalong juxing le huitan
$\mathbf{x}'$
decoder
$P(\mathbf{x}'|\mathbf{y}; \overleftarrow{\theta})$
Bush held a talk with Sharon
$\mathbf{y}$
encoder
$P(\mathbf{y}|\mathbf{x}; \overrightarrow{\theta})$
bushi yu shalong juxing le huitan
$\mathbf{x}$
(b)
Bush held a talk with Sharon
$\mathbf{y}'$
decoder
$P(\mathbf{y}'|\mathbf{x}; \overrightarrow{\theta})$
bushi yu shalong juxing le huitan
$\mathbf{x}$
encoder
$P(\mathbf{x}|\mathbf{y}; \overleftarrow{\theta})$
Bush held a talk with Sharon
$\mathbf{y}$

Fig. 3.1 Our idea is to leverage autoencoders to exploit monolingual corpora for NMT. Examples of **a** source autoencoder and **b** target autoencoder on monolingual corpora. Our idea is to leverage autoencoders to exploit monolingual corpora for NMT. In a source autoencoder, the source-to-target model $P(\mathbf{y}|\mathbf{x}; \overrightarrow{\theta})$ serves as an encoder to transform the observed source sentence $\mathbf{x}$ into a latent target sentence $\mathbf{y}$ (highlighted in grey), from which the target-to-source model $P(\mathbf{x}'|\mathbf{y}; \overleftarrow{\theta})$ reconstructs a copy of the observed source sentence $\mathbf{x}'$ from the latent target sentence. As a result, monolingual corpora can be combined with parallel corpora to train bidirectional NMT models in a semi-supervised setting

$$
\begin{aligned}
& P(\mathbf{y}'|\mathbf{y}; \overrightarrow{\theta}, \overleftarrow{\theta}) \\
= & \sum_{\mathbf{x}} P(\mathbf{y}', \mathbf{x}|\mathbf{y}; \overrightarrow{\theta}, \overleftarrow{\theta}) \\
= & \sum_{\mathbf{x}} \underbrace{P(\mathbf{x}|\mathbf{y}; \overleftarrow{\theta})}_{encoder} \underbrace{P(\mathbf{y}'|\mathbf{x}; \overrightarrow{\theta})}_{decoder}
\end{aligned}
\tag{3.2}
$$

where $\mathbf{y}$ is an observed target sentence, $\mathbf{y}'$ is a copy of $\mathbf{y}$ to be reconstructed, and $\mathbf{x}$ is a latent source sentence.

We refer to Eq. (3.2) as a *target autoencoder*. Our definition of auotoencoders is inspired by Ammar et al. [1]. We map the target sentences to the source sentences not the low dimension vector as the conventional autoencoders do. Likewise, given a monolingual corpus of source language $\mathcal{S} = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$, it is natural to introduce a *source autoencoder* that aims at reconstructing the observed source sentence via a latent target sentence:

$$
\begin{aligned}
& P(\mathbf{x}'|\mathbf{x}; \overrightarrow{\theta}, \overleftarrow{\theta}) \\
= & \sum_{\mathbf{y}} P(\mathbf{x}', \mathbf{y}|\mathbf{x}; \overleftarrow{\theta}) \\
= & \sum_{\mathbf{y}} \underbrace{P(\mathbf{y}|\mathbf{x}; \overrightarrow{\theta})}_{encoder} \underbrace{P(\mathbf{x}'|\mathbf{y}; \overleftarrow{\theta})}_{decoder}
\end{aligned}
\tag{3.3}
$$

Please see Fig. 3.1a for illustration.

3.2.3 Semi-supervised Learning

As the autoencoders involve both source-to-target and target-to-source models, it is natural to combine parallel corpora and monolingual corpora to learn bidirectional NMT translation models in a semi-supervised setting.

Formally, given a parallel corpus $\mathcal{D} = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$, a monolingual corpus of target language $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$, and a monolingual corpus of source language $\mathcal{S} = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$, we introduce our new semi-supervised training objective as follows:

$$\begin{aligned}
& J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \\
& = \underbrace{\sum_{n=1}^{N} \log P(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}; \overrightarrow{\boldsymbol{\theta}})}_{source-to-targetlikelihood} + \underbrace{\sum_{n=1}^{N} \log P(\mathbf{x}^{(n)} | \mathbf{y}^{(n)}; \overleftarrow{\boldsymbol{\theta}})}_{target-to-sourcelikelihood} \\
& + \lambda_1 \underbrace{\sum_{t=1}^{T} \log P(\mathbf{y}' | \mathbf{y}^{(t)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}_{targetautoencoder} + \lambda_2 \underbrace{\sum_{s=1}^{S} \log P(\mathbf{x}' | \mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}_{sourceautoencoder}
\end{aligned} \tag{3.4}$$

where λ_1 and λ_2 are hyper-parameters for balancing the preference between likelihood and autoencoders. Note that the objective consists of three parts: source-to-target likelihood, target-to-source likelihood, and autoencoders. λ_1 and λ_2 indicate whether we add the source or target monolingual corpus to our semi-supervised learning. If we set λ_2 to zero and λ_1 to nonzero, it means that we use the target autoencoder, and vice versa. It is also possible that these two hyper-parameters are all nonzero so we can combine both of them.

The optimal model parameters are given by

Algorithm 1 Training procedure for bidirectional neural machine translation with parallel corpora and monolingual corpora

```
procedure TrainBidrectionNMT($\mathcal{D}$, $\mathcal{T}$, $\overleftarrow{\boldsymbol{\theta}}$, $\overrightarrow{\boldsymbol{\theta}}$)
    for each $t \in [1, MaxIter]$ do
        fetch a batch of monolingual sentences $T$ from $\mathcal{T}$;
        generate n best translation candidates $\tilde{\mathcal{X}}(\mathbf{y}^{(t)})$ for each $\mathbf{y}$ in $T$ using beam search with NMT parameter $\overrightarrow{\boldsymbol{\theta}}$;
        fetch a batch of parallel sentences $D$ from $\mathcal{D}$ with equivalent number;
        calculate the gradient according to Eqs. (3.7) and (3.9);
        update the parameters $\overleftarrow{\boldsymbol{\theta}}$ and $\overrightarrow{\boldsymbol{\theta}}$;
    end for
end procedure
```

$$\begin{aligned}\overrightarrow{\boldsymbol{\theta}}^{*} = \operatorname{argmax} \Bigg\{ & \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overrightarrow{\boldsymbol{\theta}}) + \\ & \lambda_1 \sum_{t=1}^{T} \log P(\mathbf{y}'|\mathbf{y}^{(t)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) + \\ & \lambda_2 \sum_{s=1}^{S} \log P(\mathbf{x}'|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \Bigg\}\end{aligned} \tag{3.5}$$

$$\begin{aligned}\overleftarrow{\boldsymbol{\theta}}^{*} = \operatorname{argmax} \Bigg\{ & \sum_{n=1}^{N} \log P(\mathbf{x}^{(n)}|\mathbf{y}^{(n)}; \overleftarrow{\boldsymbol{\theta}}) + \\ & \lambda_1 \sum_{t=1}^{T} \log P(\mathbf{y}'|\mathbf{y}^{(t)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) + \\ & \lambda_2 \sum_{s=1}^{S} \log P(\mathbf{x}'|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}) \Bigg\}\end{aligned} \tag{3.6}$$

It is clear that the source-to-target and target-to-source models are connected via the autoencoder and can hopefully benefit each other in the joint training.

3.2.4 Training

We use mini-batch stochastic gradient descent to train our joint model. For each iteration, besides the mini-batch from the parallel corpus, we also construct another mini-batch by randomly selecting sentences from the monolingual corpus. Then, gradients are collected from both the two mini-batches to update model parameters.

The partial derivative of $J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})$ with respect to the source-to-target model $\overrightarrow{\boldsymbol{\theta}}$ is given by

$$\begin{aligned}& \frac{\partial J(\overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}} \\ & = \sum_{n=1}^{N} \frac{\partial \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overrightarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}} \\ & \quad + \lambda_1 \sum_{t=1}^{T} \frac{\partial \log P(\mathbf{y}'|\mathbf{y}^{(t)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}} \\ & \quad + \lambda_2 \sum_{s=1}^{S} \frac{\partial \log P(\mathbf{x}'|\mathbf{x}^{(s)}; \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}}\end{aligned} \tag{3.7}$$

The partial derivative with respect to $\overleftarrow{\boldsymbol{\theta}}$ can be calculated similarly.

The first term in Eq. (3.7) is same to the derivates of model likelihood. But unfortunately, the second and the last terms are intractable to calculate due to the exponential search space. As the same problem exists in the second and the last terms, we take the last term as example for showing how to effectively solve this issue. Equation (3.8) gives more detailed derivates of the last term, in which $\mathcal{X}(\mathbf{y})$ denotes the intractable search space.

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}(\mathbf{y})} P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})\frac{\partial \log P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}}}{\sum_{\mathbf{x}\in\mathcal{X}(\mathbf{y})} P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})} \tag{3.8}$$

Alternatively, we propose to use a subset of the full space $\tilde{\mathcal{X}}(\mathbf{y}) \subset \mathcal{X}(\mathbf{y})$ to approximate Eq. (3.8):

$$\frac{\sum_{\mathbf{x}\in\tilde{\mathcal{X}}(\mathbf{y})} P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})\frac{\partial \log P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}{\partial \overrightarrow{\boldsymbol{\theta}}}}{\sum_{\mathbf{x}\in\tilde{\mathcal{X}}(\mathbf{y})} P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})} \tag{3.9}$$

In practice, we use the top-k list of candidate translations of $\mathbf{y}$ as $\tilde{\mathcal{X}}(\mathbf{y})$. As $|\tilde{\mathcal{X}}(\mathbf{y}) \ll \mathcal{X}|(\mathbf{y})|$, it is possible to calculate Eq. (3.9) efficiently by enumerating all candidates in $\tilde{\mathcal{X}}(\mathbf{y})$. In practice, we find this approximation results in significant improvements and $k = 10$ seems to suffice to keep the balance between efficiency and translation quality. In addition, we think sampling method may also be a good alternative to approximate the search space which could give better diversity but a large number of candidate translations are required. In this paper, we choose the top-k list technique which sufficiently solves exponential search space problem.

3.3 Experiments

3.3.1 Setup

We evaluated our approach on the Chinese-English dataset.

As shown in Table 3.1, we use both a parallel corpus and two monolingual corpora as the training set. The parallel corpus from LDC consists of 2.56M sentence pairs with 67.53M Chinese words and 74.81M English words. The vocabulary sizes of Chinese and English are 0.21M and 0.16M, respectively. We use the Chinese and English parts of the Xinhua portion of the Gigaword corpus as the monolingual corpora. The Chinese monolingual corpus contains 18.75M sentences with 451.94M words. The English corpus contains 22.32M sentences with 399.83M words. The vocabulary sizes of Chinese and English are 0.97M and 1.34M, respectively.

For Chinese-to-English translation, we use the NIST 2006 Chinese-English dataset as the validation set for hyper-parameter optimization and model selection.

Table 3.1 Characteristics of parallel and monolingual corpora

		Chinese	English
Parallel	# Sent.	2.56M	
	# Word	67.54M	74.82M
	Vocab.	0.21M	0.16M
Monolingual	# Sent.	18.75M	22.32M
	# Word	451.94M	399.83M
	Vocab.	0.97M	1.34M

The NIST 2002, 2003, 2004, and 2005 datasets serve as test sets. Each Chinese sentence has four reference translations. For English-to-Chinese translation, we use the NIST datasets in a reverse direction: treating the first English sentence in the four reference translations as a source sentence and the original input Chinese sentence as the single reference translation. The evaluation metric is case-insensitive BLEU [16] as calculated by the `multi-bleu.perl` script.

We compared our approach with two state-of-the-art SMT and NMT systems:

1. Moses [12]: a phrase-based SMT system;
2. RNNSearch [2]: an attention-based NMT system.

For Moses, we use the default setting to train the phrase-based translation on the parallel corpus and optimize the parameters of log-linear models using the minimum error rate training algorithm [15]. We use the SRILM toolkit [19] to train 4-gram language models.

For RNNSearch, we use the parallel corpus to train the attention-based neural translation models. We set the vocabulary size of word embeddings to 30 K for both Chinese and English. We follow [14] to address rare words.

On top of RNNSearch, our approach is capable of training bidirectional attention-based neural translation models on the concatenation of parallel and monolingual corpora. The sample size k is set to 10. We set the hyper-parameter $\lambda_1 = 0.1$ and $\lambda_2 = 0$ when we add the target monolingual corpus, and $\lambda_1 = 0$ and $\lambda_2 = 0.1$ for source monolingual corpus incorporation. The threshold of gradient clipping is set to 0.05. The parameters of our model are initialized by the model trained on parallel corpus.

3.3.2 *Effect of Sample Size k*

As the inference of our approach is intractable, we propose to approximate the full search space with the top-k list of candidate translations to improve efficiency (see Eq. (3.9)).

Fig. 3.2 Effect of sample size *k* on the Chinese-to-English validation set

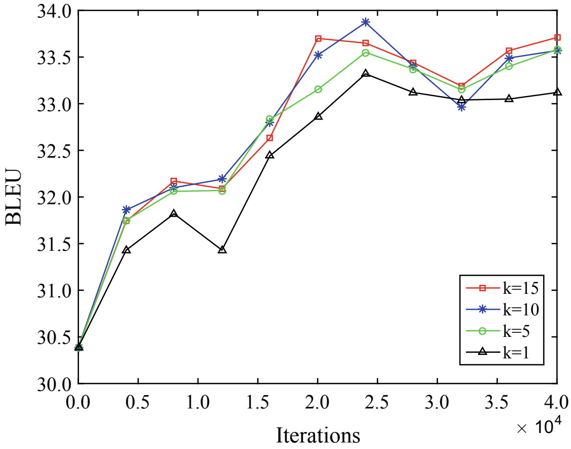


Fig. 3.3 Effect of sample size *k* on the English-to-Chinese validation set

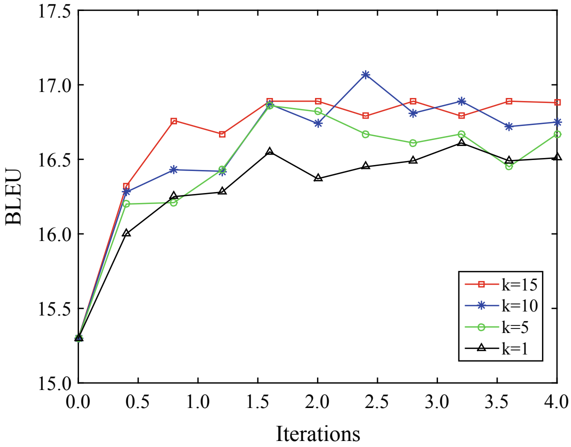


Figure 3.2 shows the BLEU scores of various settings of k over time. Only the English monolingual corpus is appended to the training data. We observe that increasing the size of the approximate search space generally leads to improved BLEU scores. There are significant gaps between $k = 1$ and $k = 5$. However, keeping increasing k does not result in significant improvements and decreases the training efficiency. We find that $k = 10$ achieves a balance between training efficiency and translation quality. As shown in Fig. 3.3, similar findings are also observed on the English-to-Chinese validation set. Therefore, we set $k = 10$ in the following experiments.

3.3.3 Effect of OOV Ratio

Given a parallel corpus, what kind of monolingual corpus is most beneficial for improving translation quality? To answer this question, we investigate the effect of *OOV ratio* on translation quality, which is defined as

$$\text{ratio} = \frac{\sum_{y \in \mathbf{y}} \| y \notin \mathcal{V}_{D_t} \|}{|\mathbf{y}|} \tag{3.10}$$

where $\mathbf{y}$ is a target-language sentence in the monolingual corpus $\mathcal{T}$, y is a target-language word in $\mathbf{y}$, $\mathcal{V}_{D_t}$ is the vocabulary of the target side of the parallel corpus D.

Intuitively, the OOV ratio indicates how a sentence in the monolingual resembles the parallel corpus. If the ratio is 0, all words in the monolingual sentence also occur in the parallel corpus.

Figure 3.4 shows the effect of OOV ratio on the Chinese-to-English validation set. Only English monolingual corpus is appended to the parallel corpus during training. We constructed four monolingual corpora of the same size in terms of sentence pairs. "0% OOV" means the OOV ratio is 0% for all sentences in the monolingual corpus. "10% OOV" suggests that the OOV ratio is no greater 10% for each sentence in the monolingual corpus. We find that using a monolingual corpus with a lower OOV ratio generally leads to higher BLEU scores. One possible reason is that low-OOV monolingual corpus is relatively easier to reconstruct than its high-OOV counterpart and results in better estimation of model parameters.

Figure 3.5 shows the effect of OOV ratio on the English-to-Chinese validation set. Only English monolingual corpus is appended to the parallel corpus during training. We find that "0% OOV" still achieves the highest BLEU scores.

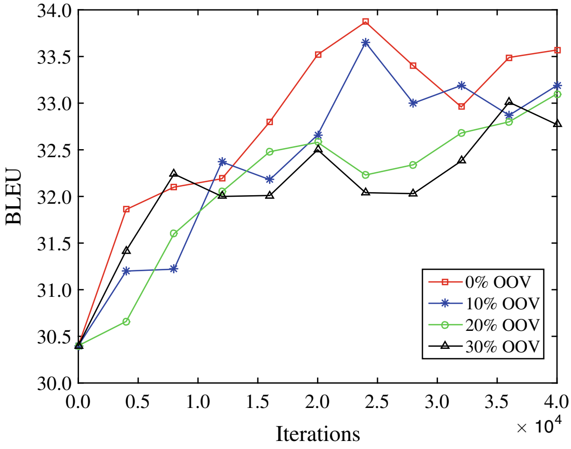


Fig. 3.4 Effect of OOV ratio on the Chinese-to-English validation set

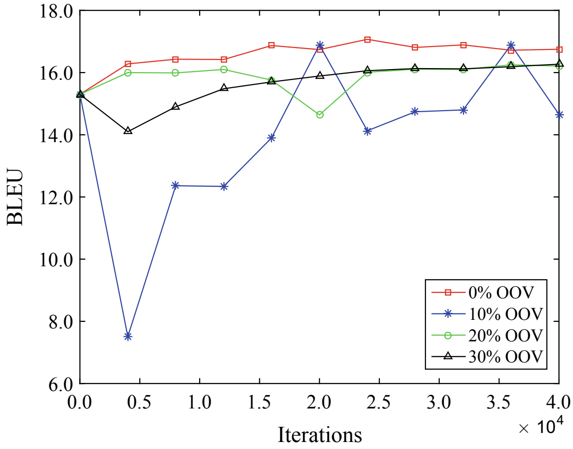


Fig. 3.5 Effect of OOV ratio on the English-to-Chinese validation set

3.3.4 Comparison with SMT

Table 3.2 shows the comparison between MOSES and our work. MOSES used the monolingual corpora as shown in Table 3.1: 18.75M Chinese sentences and 22.32M English sentences. We find that exploiting monolingual corpora dramatically improves translation performance in both Chinese-to-English and English-to-Chinese directions. Relying only on parallel corpus, RNNSEARCH outperforms MOSES trained only on parallel corpus. But the capability of making use of abundant monolingual corpora enables MOSES to achieve much higher BLEU scores than RNNSEARCH only using parallel corpus.

Instead of using all sentences in the monolingual corpora, we constructed smaller monolingual corpora with zero OOV ratio: 2.56M Chinese sentences with 47.51M words and 2.56M English sentences with 37.47M words. In other words, the monolingual corpora we used in the experiments are much smaller than those used by MOSES.

By adding English monolingual corpus, our approach achieves substantial improvements over RNNSEARCH using only parallel corpus (up to +4.7 BLEU points). In addition, significant improvements are also obtained over MOSES using both parallel and monolingual corpora (up to +3.5 BLEU points).

An interesting finding is that adding English monolingual corpora helps to improve English-to-Chinese translation over RNNSEARCH using only parallel corpus (up to +3.2 BLEU points), suggesting that our approach is capable of improving NMT using source-side monolingual corpora.

In the English-to-Chinese direction, we obtain similar findings. In particular, adding Chinese monolingual corpus leads to more benefits to English-to-Chinese translation than adding English monolingual corpus. We also tried to use both Chi-

Table 3.2 Comparison with MOSES and RNNSEARCH. MOSES is a phrase-based statistical machine translation system [12]. RNNSEARCH is an attention-based neural machine translation system [2]. "CE" donates Chinese-English parallel corpus, "C" donates Chinese monolingual corpus, and "E" donates English monolingual corpus. "√" means the corpus is included in the training data and × means not included. "NIST06" is the validation set and "NIST02-05" are test sets. The BLEU scores are case-insensitive. "*": significantly better than MOSES ($p < 0.05$); "**": significantly better than MOSES ($p < 0.01$); "+": significantly better than RNNSEARCH ($p < 0.05$); "++": significantly better than RNNSEARCH ($p < 0.01$)

System	Training data			Direction	NIST06	NIST02	NIST03	NIST04	NIST05
	CE	C	E						
MOSES	√	×	×	C→E	32.48	32.69	32.39	33.62	30.23
				E→C	14.27	18.28	15.36	13.96	14.11
	√	×	√	C→E	34.59	35.21	35.71	35.56	33.74
	√	√	×	E→C	20.69	25.85	19.76	18.77	19.74
RNNSEARCH	√	×	×	C→E	30.74	35.16	33.75	34.63	31.74
				E→C	15.71	20.76	16.56	16.85	15.14
	√	×	√	C→E	35.61^{**++}	38.78^{**++}	38.32^{**++}	38.49^{**++}	36.45^{**++}
				E→C	17.59^{++}	23.99 $^{++}$	18.95^{++}	18.85^{++}	17.91^{++}
	√	√	×	C→E	35.01^{++}	38.20^{**++}	37.99^{**++}	38.16^{**++}	36.07^{**++}
				E→C	21.12^{*++}	29.52^{**++}	20.49^{**++}	21.59^{**++}	19.97^{++}

nese and English monolingual corpora through simply setting all the λ to 0.1 but failed to obtain further significant improvements over independent addition.

Therefore, our findings can be summarized as follows:

1. Adding target monolingual corpus improves over using only parallel corpus for source-to-target translation;
2. Adding source monolingual corpus also improves over using only parallel corpus for source-to-target translation, but the improvements are smaller than adding target monolingual corpus;
3. Adding both source and target monolingual corpora does not lead to further significant improvements.

3.3.5 Comparison with Previous Work

We re-implemented Sennrich et al.'s [17] method on top of RNNSEARCH as follows:

1. Train the target-to-source neural translation model $P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$ on the parallel corpus $D = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$.
2. The trained target-to-source model $\overleftarrow{\boldsymbol{\theta}}^*$ is used to translate a target monolingual corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$ into a source monolingual corpus $\tilde{\mathcal{S}} = \{\tilde{\mathbf{x}}^{(t)}\}_{t=1}^{T}$.

Table 3.3 Comparison with Sennrich et al. [17]. Both Sennrich et al. [17] and our approach build on top of RNNSEARCH to exploit monolingual corpora. The BLEU scores are case-insensitive. “*”: significantly better than Sennrich et al. [17] ($p < 0.05$); “**”: significantly better than Sennrich et al. [17] ($p < 0.01$)

Method	Training data			Direction	NIST06	NIST02	NIST03	NIST04	NIST05
	CE	C	E						
Sennrich et al.	√	×	√	C →E	34.10	36.95	36.80	37.99	35.33
	√	√	×	E →C	19.85	28.83	20.61	20.54	19.17
This work	√	×	√	C →E	35.61**	38.78**	38.32**	38.49*	36.45**
				E →C	17.59	23.99	18.95	18.85	17.91
	√	√	×	C →E	35.01**	38.20**	37.99**	38.16	36.07**
				E →C	21.12**	29.52**	20.49	21.59**	19.97**

3. The target monolingual corpus is paired with its translations to form a pseudo parallel corpus, which is then appended to the original parallel corpus to obtain a larger parallel corpus: $\tilde{\mathcal{D}} = \mathcal{D} \cup \langle \tilde{\mathcal{S}}, \mathcal{T} \rangle$.
4. Re-train the the source-to-target neural translation model on $\tilde{\mathcal{D}}$ to obtain the final model parameters $\overrightarrow{\boldsymbol{\theta}}^*$.

Table 3.3 shows the comparison results. Both the two approaches use the same parallel and monolingual corpora. Our approach achieves significant improvements over [17] in both Chinese-to-English and English-to-Chinese directions (up to +1.8 and +1.0 BLEU points). One possible reason is that Sennrich et al. [17] only use the pesudo parallel corpus for parameter estimation for once (see Step 4 above) and the target-to-source translation model is permanent, while our approach enables source-to-target and target-to-source models to interact with each other iteratively on both parallel and monolingual corpora.

To some extent, our approach can be seen as an iterative extension of Sennrich et al.’s [17] approach: after estimating model parameters on the pseudo parallel corpus, the learned model parameters are used to produce a better pseudo parallel corpus. Table 3.4 shows example Viterbi translations on the Chinese monolingual corpus over iterations:

$$\mathbf{x}^* = \operatorname*{argmax}_{\mathbf{x}} \left\{ P(\mathbf{y}'|\mathbf{x};\ \overrightarrow{\boldsymbol{\theta}})P(\mathbf{x}|\mathbf{y};\ \overleftarrow{\boldsymbol{\theta}}) \right\} \tag{3.11}$$

We observe that the quality of Viterbi translations generally improves over time.

Table 3.4 Example translations of sentences in the monolingual corpus during semi-supervised learning. For each example, its manual translation reference is shown. We italicize some important segments in red colour. We find our approach is capable of generating better translations of the monolingual corpus over time

Monolingual	Hongsen shuo, ruguo you na jia famu gongsi dangan yishenshifa, name tamen jiang zihui qiancheng
Reference	Hongsen said, if any *logging companies* dare to defy the law, then they will *destroy their own future*
Translation	Hun sen said, if any of *those companies* dare defy the law, then they will *have their own fate*. [*iteration 0*]
	Hun sen said if any *tree felling company* dared to break the law, then they would *kill themselves*. [*iteration 40K*]
	Hun sen said if any *logging companies* dare to defy the law, they would *destroy the future themselves*. [*iteration 240K*]
Monolingual	Dan youyu youke shuliang jizeng, renlei huodong dui senlin gongyuan de fumian yingxiang riqu xianxian
Reference	However, because of the rapid growth in the number of visitors, the negative impact of human activity on the forest park has become increasingly apparent
Translation	But because of the rapid growth in visitor numbers, the negative impact for human activity has become increasingly obvious. [*iteration 0*]
	However, because of the rapid growth in visitor numbers, the negative impact of human activity has become increasingly evident. [*iteration 40K*]
	However, because of the rapid growth in visitor numbers, the negative impact of human activity *on the park* has become increasingly apparent. [*iteration 240K*]
Monolingual	Dan yidan panjue jieguo zuizhong queding, ze bixu zai 30 tian nei zhixing
Reference	But once *the final verdict is confirmed*, it must be executed within 30 days
Translation	However, *in the final analysis*, it must be carried out within 30 days. [*iteration 0*]
	However, *in the final analysis*, the final decision will be carried out within 30 days. [*iteration 40K*]
	However, once *the verdict is finally confirmed*, it must be carried out within 30 days. [*iteration 240K*]

3.4 Summary

We have presented a semi-supervised approach to training bidirectional neural machine translation models. The central idea is to introduce autoencoders on the monolingual corpora with source-to-target and target-to-source translation models as encoders and decoders. Experiments on Chinese-English NIST datasets show that our approach leads to significant improvements.

As our method is sensitive to the OOVs present in monolingual corpora, we plan to integrate Jean et al.'s [9] technique on using very large vocabulary into our approach. It is also necessary to further validate the effectiveness of our approach on more language pairs and NMT architectures. Another interesting thing is that we can share some parameters, like word embedding, because we use two translation models, source-to-target and target-to-source models, where their vocabularies remain the same. We believe sharing parameters enables our models to better benefit each other.

References

1. Ammar, W., Dyer, C., & Smith, N. (2014). Conditional random field autoencoders for unsupervised structured prediction. In *Advances in neural information processing systems (NIPS)*.
2. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
3. Bertoldi, N., & Federico, M. (2009). Domain adaptation for statistical machine translation. In *Workshop on statistical machine translation*.
4. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.
5. Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Association for computational linguistics (ACL)*.
6. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empirical methods in natural language processing (EMNLP)*.
7. Gulccehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., et al. (2015). On using monolingual corpora in neural machine translation. arXiv:1503.03535 [cs.CL].
8. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.
9. Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Association for computational linguistics (ACL)*.
10. Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In *Empirical methods in natural language processing (EMNLP)*.
11. Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems (NIPS)*.
12. Koehn, P., & Hoang, H. (2007). Factored translation models. In *Empirical methods in natural language processing (EMNLP)*.
13. Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *North American association for computational linguistics (NAACL)*.
14. Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Association for computational linguistics (ACL)*.
15. Och, F. (2003). Minimum error rate training in statistical machine translation. In *Association for computational linguistics (ACL)*.
16. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Association for computational linguistics (ACL)*.
17. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Association for computational linguistics (ACL)*.
18. Socher, R., Pennington, J., Huang, E., Ng, A., & Manning, C. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Empirical methods in natural language processing (EMNLP)*.
19. Stolcke, A. (2002). SRILM—an extensible language modeling toolkit. In *International conference on spoken language processing*.

20. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*.
21. Ueffing, N., Haffari, G., & Sarkar, A. (2007). Trasductive learning for statistical machine translation. In *Association for computational linguistics (ACL)*.
22. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*.

Chapter 4
Joint Training for Pivot-Based Neural Machine Translation

Abstract While recent neural machine translation approaches have delivered state-of-the-art performance for resource-rich language pairs, they suffer from the data scarcity problem for resource-scarce language pairs. Although this problem can be alleviated by exploiting a pivot language to bridge the source and target languages, the source-to-pivot and pivot-to-target translation models are usually independently trained. In this work, we introduce a joint training algorithm for pivot-based neural machine translation. We propose three methods to connect the two models and enable them to interact with each other during training. Experiments on Europarl and WMT corpora show that joint training of source-to-pivot and pivot-to-target models leads to significant improvements over independent training across various languages.

4.1 Introduction

Recent several years have witnessed the rapid development of neural machine translation (NMT) [1, 15], which advocates the use of neural networks to directly model the translation process in an end-to-end way. Thanks to the capability of learning representations from training data, NMT systems have achieved significant improvements over conventional statistical machine translation (SMT) across a variety of language pairs [9, 10].

However, there still remains a major challenge for NMT: large-scale parallel corpora are usually non-existent for most language pairs. This is unfortunate because NMT is a data-hungry approach and requires a large amount of data to fully train model parameters. Without sufficient training data, NMT tends to learn poor estimates on low-count events. Zoph et al. [20] indicate that NMT obtains much worse translation quality than SMT when only small-scale parallel corpora are available.

As a result, improving neural machine translation on resource-scarce language pairs has attracted much attention in the community [6, 9, 20]. Most existing methods focus on leveraging data of multiple resource-rich language pairs to improve NMT for resource-scarce language pairs. Firat et al. [6] propose multi-way, multilingual neural machine translation to achieve direct source-to-target translation even without parallel data available. Zoph et al. [20] present a transfer learning method

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_4

that transfers the model parameters trained for resource-rich language pairs to initialize and constrain the translation model training of resource-scarce language pairs. Johnson et al. [9] introduce a universal NMT model for all language pairs, which takes advantage of multilingual data to improve NMT for all languages involved.

Bridging source and target languages with a *pivot* language is another important direction, which has been intensively studied in conventional SMT [2, 4, 5, 16–18]. Pivot-based approaches assume that there exist source-pivot and pivot-target parallel corpora, which can be used to train source-to-pivot and pivot-to-target translation models, respectively. One of the most representative approaches, triangulation approach, is to construct a source-to-target phrase table through combining source-to-pivot and pivot-to-target phrase tables. Another representative approach adopts a pivot-based translation strategy. As a result, source-to-target translation can be divided into two steps: the source sentence is first translated into a pivot sentence using the source-to-pivot model, which is then translated to a target sentence using the pivot-to-target model. Pivot-based approaches have been widely used in SMT due to its simplicity, effectiveness, and minimum requirement of multilingual data. Recently, Johnson et al. [8] adapt pivot-based approaches to NMT and show that their universal model without incremental training achieves much worse translation performance than pivot-based NMT.

However, pivot-based approaches often suffer from the error propagation problem: the errors made in the source-to-pivot translation will be propagated to the pivot-to-target translation. This can be partly attributed to the discrepancy between source-pivot and pivot-target parallel corpora since they are usually loosely-related or even unrelated. To aggregate the situation, source-to-pivot and pivot-to-target translation models are trained independently, which further enlarges the gap between source and target languages.

In this work, we propose an approach to joint training for pivot-based neural machine translation. The basic idea is to connect the source-to-pivot and pivot-to-target NMT models and enable them to interact with each other during training. This can be done either by encouraging the sharing of word embeddings on the pivot language or by maximizing the likelihood of the cascaded model on a small source-target parallel corpus. Experiments on the Europarl and WMT corpora show that joint training of source-to-pivot and pivot-to-target models obtains significant improvements over independent training.

4.2 Pivot-Based NMT

Given a source language sentence $\mathbf{x}$ and a target language sentence $\mathbf{y}$, we use $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ to denote a standard attention-based neural machine translation model [1], where $\boldsymbol{\theta}_{x \to y}$ is a set of model parameters.

Ideally, the source-to-target model can be trained on a source-target parallel corpus $D_{x,y} = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$ using maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}_{x\rightarrow y} = \underset{\boldsymbol{\theta}_{x\rightarrow y}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}_{x\rightarrow y}) \right\} \tag{4.1}$$

where the log-likelihood is defined as

$$\mathcal{L}(\boldsymbol{\theta}_{x\rightarrow y}) = \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \boldsymbol{\theta}_{x\rightarrow y}) \tag{4.2}$$

Unfortunately, parallel corpora are usually not readily available for low-resource language pairs. Instead, one can assume that there exist a third language called *pivot* with source-pivot and pivot-target parallel corpora available. As a result, it is possible to bridge the source and target languages with the pivot [2, 4, 5, 16–18]. There usually exist a large number of parallel corpora between $\mathcal{S}$ or $\mathcal{T}$ and $\mathcal{P}$ despite no direct parallel corpora between $\mathcal{S}$ and $\mathcal{T}$. $\mathcal{P}$ can serve as the pivot language to "bridge" the translation between $\mathcal{S}$ and $\mathcal{T}$. In statical machine translation, the *triangulation* approach [4, 17] is one of the most representative approaches. Given a source-to-pivot phrase table $\mathcal{M}_{sp}$ and a pivot-to-target phrase table $\mathcal{M}_{pt}$, the source-to-target phrase table $\mathcal{M}_{st}$ is obtained by merging $\mathcal{M}_{sp}$ and $\mathcal{M}_{pt}$:

$$\mathcal{M}_{st} = \mathcal{M}_{sp} \otimes \mathcal{M}_{pt} \tag{4.3}$$

where $\otimes$ is merging operation, such as multiplication of translation probability of source-to-pivot and pivot-to-target phrases with identical pivot-language phrases.

The constructed source-to-target phrase table $\mathcal{M}_{st}$ facilitates building a source-to-target statistical machine translation system. However, free of latent structures such as phrase table, the NMT model directly maximizes the conditional probability $p(\mathbf{y}|\mathbf{x})$ as Eq. (4.2) indicates. Therefore, it is non-trivial to utilize pivot languages in NMT.

Let $\mathbf{z}$ be a pivot language sentence. The source-to-target model can be decomposed into two sub-models by treating the pivot sentence as a latent variable:

$$\begin{aligned} & P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y}) \\ = & \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y}) \end{aligned} \tag{4.4}$$

Let $D_{x,z} = \{\langle \mathbf{x}^{(m)}, \mathbf{z}^{(m)} \rangle\}_{m=1}^{M}$ be a source-pivot parallel corpus, and $D_{z,y} = \{\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$ be a pivot-target parallel corpus. The source-to-pivot and pivot-to-target models can be **independently** trained on the two parallel corpora, respectively:

$$\hat{\boldsymbol{\theta}}_{x\rightarrow z} = \underset{\boldsymbol{\theta}_{x\rightarrow z}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}_{x\rightarrow z}) \right\} \tag{4.5}$$

$$\hat{\boldsymbol{\theta}}_{z\rightarrow y} = \underset{\boldsymbol{\theta}_{z\rightarrow y}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}_{z\rightarrow y}) \right\} \tag{4.6}$$

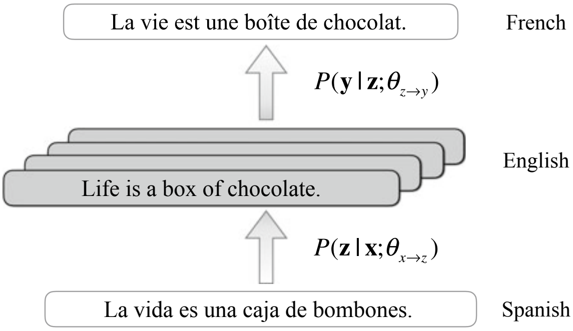

La vie est une boîte de chocolat.
French
$P(\mathbf{y} \mid \mathbf{z}; \theta_{z \to y})$
English
Life is a box of chocolate.
$P(\mathbf{z} \mid \mathbf{x}; \theta_{x \to z})$
La vida es una caja de bombones.
Spanish

Fig. 4.1 The illustration of translation on Spanish-French with English as the pivot language. The Spanish-English NMT model $P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\to z})$ first transforms a Spanish sentence into latent English sentences, from which English-French NMT model $P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\to y})$ attempts to generate a French sentence corresponding to the Spanish sentence

where the log-likelihoods are defined as:

$$\mathcal{L}(\boldsymbol{\theta}_{x\to z}) = \sum_{m=1}^{M} \log P(\mathbf{z}^{(m)}|\mathbf{x}^{(m)}; \boldsymbol{\theta}_{x\to z}) \tag{4.7}$$

$$\mathcal{L}(\boldsymbol{\theta}_{z\to y}) = \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{z}^{(n)}; \boldsymbol{\theta}_{z\to y}) \tag{4.8}$$

As Fig. 4.1 shows, a pivot-based translation strategy is usually adopted. Given an unseen source sentence to be translated $\mathbf{x}$, the decision rule is given by:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x\to z}) P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z\to y}) \right\} \tag{4.9}$$

Due to the exponential search space of the pivot language, the decoding process is usually approximated with two steps. The first step translates the source sentence $\mathbf{x}$ into a pivot sentence:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \left\{ P(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x\to z}) \right\} \tag{4.10}$$

Then, the pivot sentence is translated to a target sentence:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}_{z\to y}) \right\} \tag{4.11}$$

Although pivot-based approaches are widely for addressing the data scarcity problem in machine translation, they suffer from cascaded translation errors: the mistakes made in the source-to-pivot translation as shown in Eq. (4.10) will be propagated to the pivot-to-target translation as shown in Eq. (4.11). This can be partly attributed to the **model discrepancy** problem: the source-to-pivot and pivot-to-target models are quite different in terms of vocabulary and parameter space because the source-pivot and pivot-target parallel corpora are usually loosely-related or even unrelated. To make things worse, the source-to-pivot model $P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z})$ and the pivot-to-target model $P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y})$ are trained on the two parallel corpora **independently**, which further increases the discrepancy between two models.

Therefore, it is important to reduce the discrepancy between source-to-pivot and pivot-to-target models to further improve pivot-based neural machine translation.

4.3 Joint Training for Pivot-Based NMT

4.3.1 Training Objective

To alleviate the model discrepancy problem, we propose an approach to joint training for pivot-based neural machine translation. The basic idea is to connect source-to-pivot and pivot-to-target models and enable them to interact with each other during training. Our new training objective is given by:

$$\begin{aligned}&\mathcal{J}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})\\ &= \mathcal{L}(\boldsymbol{\theta}_{x\rightarrow z}) + \mathcal{L}(\boldsymbol{\theta}_{z\rightarrow y}) + \lambda\mathcal{R}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})\end{aligned} \tag{4.12}$$

Note that the training objective consists of three parts: the source-to-pivot likelihood $\mathcal{L}(\boldsymbol{\theta}_{x\rightarrow z})$, the pivot-to-target likelihood $\mathcal{L}(\boldsymbol{\theta}_{z\rightarrow y})$, and a connection term $\mathcal{R}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$. The hyper-parameter λ is used to balance the preference between likelihoods and the connection term.

We expect that the connection term associates the source-to-pivot model $\boldsymbol{\theta}_{x\rightarrow z}$ with the pivot-to-target model $\boldsymbol{\theta}_{z\rightarrow y}$ and enables the interaction between two models during training. In the following subsection, we will introduce the three connection terms used in our experiments.

4.3.2 Connection Terms

It is difficult to connect the source-to-pivot and pivot-to-target models during training because the source-to-pivot and pivot-to-target models are distantly-related by definition. More importantly, NMT lacks linguistically interpretable language structures such as phrases in SMT to achieve a direct connection at the parameter level [17].

Fortunately, both the source-to-pivot and pivot-to-target models include the word embeddings of the pivot language as parameters. It is possible to connect the two models via pivot word embeddings.

More formally, let $\mathcal{V}^z_{x\to z}$ be the pivot vocabulary of the source-to-pivot model and $\mathcal{V}^z_{z\to y}$ be the pivot vocabulary of the pivot-to-target model. We use w to denote a word in the pivot language and $\boldsymbol{\theta}^w_{x\to z} \in \mathbb{R}^d$ to denote the vector representation of w in the source-to-pivot model. $\boldsymbol{\theta}^w_{z\to y} \in \mathbb{R}^d$ is defined in a similar way.

Our first connection term encourages the two models to generate the same vector representations for pivot words in the intersection of two vocabularies:

$$\begin{aligned} &\mathcal{R}_{\text{hard}}(\boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}) \\ &= \prod_{w \in \mathcal{V}^z_{x\to z} \cap \mathcal{V}^z_{z\to y}} \delta(\boldsymbol{\theta}^w_{x\to z}, \boldsymbol{\theta}^w_{z\to y}) \end{aligned} \tag{4.13}$$

where $\delta(\boldsymbol{\theta}^w_{x\to z}, \boldsymbol{\theta}^w_{z\to y}) = 1$ if the two vectors $\boldsymbol{\theta}^w_{x\to z}$ and $\boldsymbol{\theta}^w_{z\to y}$ are identical. Otherwise, $\delta(\boldsymbol{\theta}^w_{x\to z}, \boldsymbol{\theta}^w_{z\to y}) = 0$.

As word embeddings seem hardly to be exactly identical due to the divergence of natural languages, an alternative is to soften the above hard matching constraint by penalizing the Euclidean distance between two vectors:

$$\begin{aligned} &\mathcal{R}_{\text{soft}}(\boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}) \\ &= - \sum_{w \in \mathcal{V}^z_{x\to z} \cap \mathcal{V}^z_{z\to y}} ||\boldsymbol{\theta}^w_{x\to z} - \boldsymbol{\theta}^w_{z\to y}||_2 \end{aligned} \tag{4.14}$$

The third connection term assumes that there is a small bridging source-target parallel corpus $D_{x,y} = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^S$ (*Bridging Corpus*) available. The connection term is defined as the log-likelihood of the bridging data:

$$\begin{aligned} &\mathcal{R}_{\text{likelihood}}(\boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}) \\ &= \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}) && (4.15) \\ &= \sum_{s=1}^{S} \log \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}^{(s)}; \boldsymbol{\theta}_{x\to z}) P(\mathbf{y}^{(s)}|\mathbf{z}; \boldsymbol{\theta}_{z\to y}) && (4.16) \end{aligned}$$

4.3.3 Training

In training, our goal is to find the optimal source-to-pivot and pivot-to-target model parameters that maximize the training objective:

$$\hat{\boldsymbol{\theta}}_{x\to z}, \hat{\boldsymbol{\theta}}_{z\to y} = \mathop{\text{argmax}}_{\boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}} \left\{ \mathcal{J}(\boldsymbol{\theta}_{x\to z}, \boldsymbol{\theta}_{z\to y}) \right\} \tag{4.17}$$

The partial derivative of $\mathcal{J}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$ with respect to the parameters $\boldsymbol{\theta}_{x\rightarrow z}$ of the source-to-pivot model can be calculated as:

$$\begin{aligned} &\frac{\partial \mathcal{J}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})}{\partial \boldsymbol{\theta}_{x\rightarrow z}} \\ &= \sum_{m=1}^{M} \frac{\partial \log P(\mathbf{z}^{(m)}|\mathbf{x}^{(m)}; \boldsymbol{\theta}_{x\rightarrow z})}{\partial \boldsymbol{\theta}_{x\rightarrow z}} + \lambda \frac{\partial \mathcal{R}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})}{\partial \boldsymbol{\theta}_{x\rightarrow z}} \end{aligned} \tag{4.18}$$

The partial derivative with respect to the parameters $\boldsymbol{\theta}_{z\rightarrow y}$ can be calculated similarly.

When training our joint objective, the gradients of the first and second connection terms, $\mathcal{R}_{\text{hard}}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$ and $\mathcal{R}_{\text{soft}}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$, with respect to model parameters are easy. However, calculating the gradients of the third connection term $\mathcal{R}_{\text{likelihood}}(\boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$ involves enumerating all possible pivot sentences in an exponential search space (see Eq. (4.19)).

$$\frac{\sum_{\mathbf{z}\in\mathcal{Z}(\mathbf{x})} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y}) \frac{\partial \log P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z})}{\partial \boldsymbol{\theta}_{x\rightarrow z}}}{\sum_{\mathbf{z}\in\mathcal{Z}(\mathbf{x})} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y})} \tag{4.19}$$

To alleviate this problem, we follow standard practice to use a subset $\tilde{\mathcal{Z}}(\mathbf{x}) \in \mathcal{Z}(\mathbf{x})$ to approximate the full space [3, 14]. Two methods can be used to generate a subset: sampling k translations from the full space [14] or generating a top-k list of candidate translations [3]. We find that using top-k lists leads to better results than sampling in our experiments. In experiments, we find that $k = 10$ is an advisable choice considering the efficiency and translation quality.

$$\frac{\sum_{\mathbf{z}\in\tilde{\mathcal{Z}}(\mathbf{x})} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y}) \frac{\partial \log P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z})}{\partial \boldsymbol{\theta}_{x\rightarrow z}}}{\sum_{\mathbf{z}\in\tilde{\mathcal{Z}}(\mathbf{x})} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z\rightarrow y})} \tag{4.20}$$

We use standard mini-batched stochastic gradient descent algorithms to optimize model parameters. In each iteration, three mini-batches are constructed by randomly selecting sentence pairs from the source-pivot parallel corpus $D_{x,z}$, the pivot-target parallel corpus $D_{z,y}$, and the bridging source-target parallel corpus $D_{x,y}$ (only available for the third connection term), respectively. After separate gradient calculation in each mini-batch, the gradients are collected to update model parameters.

4.4 Experiments

4.4.1 Setup

We evaluated our approach on two translation tasks:

1. Spanish-English-French: Spanish as the source language, English as the pivot language, and French as the target language,
2. German-English-French: German as the source language, English as the pivot language, and French as the target language.

Table 4.1 shows the statistics of the Europarl and WMT corpora used in our experiments. We use `tokenize.perl` script for tokenization. For each language pair, we remove the empty lines and retain sentence pairs with no more than 50 words. To avoid the intersection of the source-pivot and pivot-target corpora, we split the overlapped pivot-language sentences of source-to-pivot and pivot-to-target corpora into two separate parts with equal size and merge them separately with the non-overlapping parts for each language pair.

The Europarl corpus consists of 850K Spanish-English sentence pairs with 22.32M Spanish words and 21.44M English words, 840K German-English sentence pairs with 20.88M German words and 21.91M English words, and 900K English-French sentence pairs with 22.56M English words and 25.00M French words. The WMT 2006 shared task datasets are used as the development and test sets. The evalua-

Table 4.1 Characteristics of Spanish-English, German-English and English-French datasets on the Europarl and WMT corpora. "es" denotes Spanish, "en" denotes English, "de" denotes German, and "fr" denotes French

Corpus	Lang.		Source	Target
Europarl	es-en	# Sent.	850K	
		# Word	22.32M	21.44M
		Vocab.	118.81K	78.37K
	de-en	# Sent.	840K	
		# Word	20.88M	21.91M
		Vocab.	242.87K	80.44KM
	en-fr	# Sent.	900K	
		# Word	22.56M	25.00M
		Vocab.	80.08K	98.50K
WMT	es-en	# Sent.	6.78M	
		# Word	183.01M	166.28M
		Vocab.	0.98M	0.91M
	en-fr	# Sent.	9.29M	
		# Word	227.06M	258.95M
		Vocab.	0.23M	1.19M

tion metric is case-insensitive BLEU [12] as calculated by the `multi-bleu.perl` script.

The WMT corpus is composed of the Common Crawl, News Commentary, Europarl v7 and UN corpora. The Spanish-English parallel corpus consists of 6.78M sentence pairs with 183.01M Spanish words and 166.28M English words. The English-French parallel corpus comprises 9.29M sentence pairs with 227.06M English words and 258.95M French words. The *newstest2011* and *newstest2012* datasets serve as development and test sets. We use case-sensitive BLEU as the evaluation metric.

We use the attention-based neural machine translation system RNNSEARCH [1] in our experiments. For the Europarl corpus in Table 4.1, we set the vocabulary size of all the languages to 30K which covers over 99% of words for English, Spanish and French and over 97% for German. We follow [7] to address rare words. For Spanish-English and English-French corpora from the WMT corpus, due to large vocabulary size, we adopt byte pair encoding [13] to split rare words into sub-words. The size of sub-words is set to 43K, 33K, 43K respectively for Spanish, English, and French. These sub-words cover 100% of the text.

We set the hyper-parameter λ for balancing between likelihood and the connection term to 1.0. The threshold of gradients is set to 0.1. The bridging source-target parallel corpus contains 100K sentence pairs that do not overlap with the training data. We set k to 10 for calculating top-k lists to approximate the full search space. The parameters for the source-to-pivot and pivot-to-target translation models in the likelihood connection term are initialized by pre-trained model parameters.

4.4.2 Results on the Europarl Corpus

Table 4.2 shows the comparison results between our joint training on three connection terms and independent training on the Europarl Corpus. For the source-to-target translation task, we present source-to-pivot, pivot-to-target and source-to-target translation results compared with independent training. In Spanish-to-French translation task, soft connection achieves significant improvements in Spanish-to-French and Spanish-to-English directions although hard connection still performs comparably with independent training. In German-to-French translation task, soft and hard connections also achieve comparable performances with independent training.

In contrast, we find that likelihood connection dramatically improves translation performance on both Spanish-to-French and German-to-French corpora (up to +2.80 BLEU scores in Spanish-to-French and up to 2.23 BLEU scores in German-to-French). The significant improvements for source-to-pivot and pivot-to-target directions are also observed. This suggests that in the third connection term introducing source-to-target parallel corpus to maximize $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow z}, \boldsymbol{\theta}_{z\rightarrow y})$ with $\mathbf{z}$ as latent variables makes the source-to-pivot and pivot-to-target translation models improved collaboratively.

Table 4.2 Comparison between independent and joint training on Spanish-French and German-French translation tasks using the Europarl corpus. English is treated as the pivot language. We propose three kinds of connection terms that jointly train source-to-pivot and pivot-to-target translation models. The comparison of translation quality for source-to-pivot, pivot-to-target and source-to-target directions are shown. Source-to-target translation results are obtained by translating pivot language sentences. The BLEU scores are case-insensitive. "*": significantly better than independent training ($p < 0.05$); "**": significantly better than independent training ($p < 0.01$). We use the statistical significance test with paired bootstrap resampling [11]

Training	Connection	Dataset	Spanish-French			German-French		
			es → en	en → fr	es → fr	de → en	en → fr	de → fr
Indep.	–	Dev.	31.53	30.46	29.52	26.52	30.46	23.67
		Test	31.54	31.42	29.79	26.47	31.42	23.70
Joint	Hard	Dev.	31.81	30.18	29.11	26.48	30.47	23.87
		Test	31.55	31.13	29.93	26.58	31.35	23.88
	Soft	Dev.	32.11**	30.41	30.24**	26.92	30.39	23.99
		Test	31.96*	31.40	30.57**	26.55	31.33	23.79
	Likelihood	Dev.	33.35**	31.63**	32.45**	27.90**	31.49**	25.21**
		Test	33.54**	32.33**	32.59**	28.01**	32.34**	25.93**

Table 4.3 shows pivot and target translation examples of independent training and our approaches. Apparently, our approaches improve translation quality of both pivot sentences and target sentences.

According to Eq. (4.4), the cost of the source-to-target model can be decomposed into the cost of source-to-pivot and pivot-to-target models. Because we have a small test trilingual corpus, (Spanish, English, French), we use the English sentence to approximate the latent variables in Eq. (4.4). Then we calculate the cost of Spanish-to-French on the trilingual corpus. Figure 4.2 shows the learning curves of the test cost of independent training and joint training on three connection terms. We can find that hard and soft connections learn slower than the independent training. Likelihood connection drives its cost lower after fine-tuning based on pre-trained parameters in just 10K iterations.

4.4.3 Results on the WMT Corpus

Likelihood connection obtains the best performance in our three proposed connection terms according to experiments on the Europarl corpus. To further verify its practicability, Table 4.4 shows results on the WMT corpus which is a much larger corpus. We find that likelihood connection still outperforms independent training significantly on Spanish-to-English, English-to-French and Spanish-to-French directions (up to +1.18 BLEU scores in Spanish-to-French).

Table 4.3 Examples of pivot and target translations using the pivot-based translation strategy. We observe that our approaches generate better translations for both pivot and target sentences. We italicize *correct translation segments* which are no short than 2-grams

Ground truth	Source	uno no debe empezar a dudar en público del valor, tampoco del valor inmediato en el aspecto material, de esta ampliación
	Pivot	It makes little sense to start to doubt in public the value, including the direct value at a material level, of this enlargement
	Target	il ne faut pas commencer à douter en public de la valeur, ni de la valeur immédiate, de la portée matérielle de cet élargissement
Indep.	Pivot	One should not begin *to doubt in* terms of *the value* of courage, or of the immediate effect on material, of *enlargement.* [BLEU: 13.33]
	Target	*il ne* faudrait pas se tromper en termes de valeur de courage ou d’ effet immédiat sur le matériel, l’ *élargissement.* [BLEU: 8.69]
Hard	Pivot	One must not *start to doubt in* the public, not the immediate value in the material, *this enlargement.* [BLEU: 19.02]
	Target	*il ne faut pas* que l’ on *commence à douter*, ni au public, ni à la *valeur immédiate,* à l’ *élargissement.* [BLEU: 25.36]
Soft	Pivot	One cannot start thinking of the value of *the value,* and the immediate courage *, of this enlargement.* [BLEU: 21.57]
	Target	On *ne peut pas commencer à* penser à la valeur *de la valeur*, au courage immédiat, *de cet élargissement.* [BLEU: 26.60]
Likelihood	Pivot	One must not *start to* question the value of *the value,* either of the immediate value in the material aspect *, of this enlargement.* [BLEU: 24.60]
	Target	il ne faut pas commencer à remettre en question la valeur *de la valeur, ni de la valeur immédiate* de l’ aspect matériel, *de cet élargissement.* [BLEU: 56.40]

We also compare our approach with Firat et al. [6]. They propose a multi-way, multilingual NMT model to build a source-to-target translation model. Although our parallel training corpus is much smaller than theirs, Table 4.5 shows that our approach achieves substantial improvements over them (up to +4.32 BLEU).

Fig. 4.2 Learning curves of independent training and joint training on different connection terms

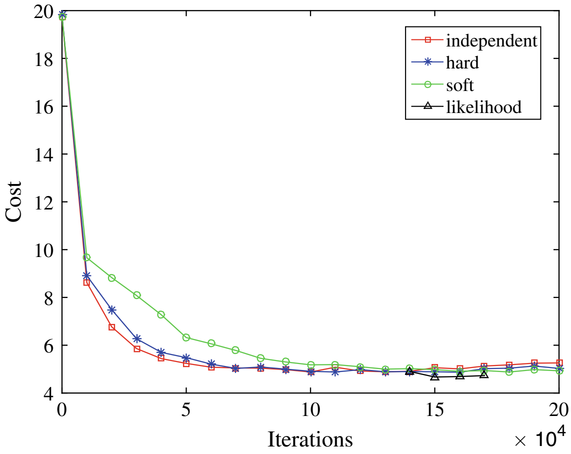


Table 4.4 Results on Spanish-French translation task from WMT corpus. English is treated as the pivot language. "**": significantly better than independent training ($p < 0.01$)

Method	Dataset	Spanish-French (WMT)		
		es → en	en → fr	es → fr
Indep.	Dev.	27.62	27.90	24.92
	Test	29.03	25.82	24.60
Likelihood	Dev.	28.92**	28.52**	26.24**
	Test	30.43**	26.36**	25.78**

Table 4.5 Comparison with Firat et al. [6]. "**": significantly better than independent training ($p < 0.01$)

Systems	*newstest2012*	*newstest2013*
Firat et al. [6]	21.81	21.46
This work	25.95**	25.78**

4.4.4 Effect of Bridging Corpora

As bridging corpora are used in likelihood connection term for "bridging" the source-to-pivot and pivot-to-target translation models, why do not we directly build NMT systems with these corpora?

We train source-to-target models using bridging corpora and show translation results in Table 4.6. We observe that performance is much worse than that in Tables 4.2 and 4.4 using the pivot-based translation strategy. It indicates that NMT yields poor performance on low-resource languages and the pivot-based translation strategy remedies the drawback to alleviate data scarcity effectively.

Table 4.6 Translation performance on bridging corpora

Corpus	Lang.	Source-target	Source-pivot-target
Europarl	es → fr	26.37	29.79
	de → fr	14.02	23.70
WMT	es → fr	11.75	24.60

Table 4.7 Effect of the data size of source-to-target parallel corpora (*Bridge Corpora*) used in LIKELIHOOD

# Sent.	es → en	en → fr	es → fr
0	31.53	30.46	29.52
1K	32.64	30.29	30.23
10K	32.92	30.93	31.51
50K	33.29	31.57	32.40
100K	33.35	31.63	32.45

We also investigate the effect of the data size of bridging corpora on the likelihood connection. Table 4.7 shows that using a small parallel corpus (1K sentence pairs) has made a measurable improvement. When more than 50K sentence pairs are added, the further improvements become modest. This finding suggests that a small corpus suffices to enable the likelihood connection to reach the reasonable performance.

4.5 Summary

We present joint training for pivot-based neural machine translation. The connection terms in our joint training objective make the source-to-pivot and pivot-to-target translation models interact with each other. Experiments on different language pairs confirm that our approach achieves significant improvements. It is appealing to combine source and pivot sentences for decoding target sentences [6] or train a multi-source model directly [19]. We also plan to study better connection terms for our joint training.

References

1. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
2. Bertoldi, N., Barbaiani, M., Federico, M., & Cattoni, R. (2008). Phrase-based statistical machine translation with pivot languages. In *International workshop on spoken language translation*.

3. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., et al. (2016). Semi-supervised learning for neural machine translation. In *Association for computational linguistics (ACL)*.
4. Cohn, T., & Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Association for computational linguistics (ACL)*.
5. El Kholy, A., Habash, N., Leusch, G., Matusov, E., & Sawaf, H. (2013). Language independent connectivity strength features for phrase pivot statistical machine translation. In *Association for computational linguistics (ACL)*.
6. Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F. T. Y., & Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. In *Empirical methods in natural language processing (EMNLP)*.
7. Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Association for computational linguistics (ACL)*.
8. Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using LSTM for region embeddings. arXiv:1602.02373.
9. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., et al. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. arXiv:1611.04558.
10. Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. arXiv:1610.01108v2.
11. Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Empirical methods in natural language processing (EMNLP)*.
12. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Association for computational linguistics (ACL)*.
13. Sennrich, R., Haddow, B., Birch, A. (2016). Neural machine translation of rare words with subword units. In *Association for computational linguistics (ACL)*.
14. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., et al. (2016). Minimum risk training for neural machine translation. In *Association for computational linguistics (ACL)*.
15. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*.
16. Utiyama, M., & Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *North American association for computational linguistics (NAACL)*.
17. Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*.
18. Zahabi, S. T., Bakhshaei, S., & Khadivi, S. (2013). Using context vectors in improving a machine translation system with bridge language. In *Association for computational linguistics (ACL)*.
19. Zoph, B., & Knight, K. (2016). Multi-source neural translation. In *North American association for computational linguistics (NAACL)*.
20. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Empirical methods in natural language processing (EMNLP)*.

Chapter 5
Joint Modeling for Bidirectional Neural Machine Translation with Contrastive Learning

Abstract Standard neural machine translation (NMT) only captures unidirectional dependencies to model a translation process from source to target. Nevertheless, the inverse information is explicitly available to reinforce the confidence of the translation process. We therefore propose an end-to-end bidirectional NMT model to associate the standard source-to-target and target-to-source translation models through a novel soft connection approach, which opens up the information interaction between two directional models. We also adopt a contrastive learning approach to train the bidirectional NMT model so as to enhance the information interaction. Experiments on Chinese-English and English-German translation tasks demonstrate that our approach significantly outperforms strong NMT and SMT baseline systems.

5.1 Introduction

Recently, neural machine translation (NMT) has achieved great success in both academia and industry [1, 26, 28]. The common architecture of NMT is based on an encoder-decoder framework. The encoder, which can be either a recurrent neural network (RNN) [1] or a convolutional neural network (CNN) [5, 6], transforms the source sentence into a sequence of continuous vector representations. The other RNN serves as the decoder to generate the target sentence based on the representations. The attention mechanisms [1] bring substantial improvements by enabling the encoder-decoder NMT to capture long-distance dependencies.

Compared with statistical machine translation (SMT) [4, 14], NMT has obtained competitive quality improvements, especially in the human evaluation [28]. However, most existing NMT systems only model the parallel sentence pairs from a single direction. They train the source-to-target and target-to-source translation models independently with no information sharing. In SMT, bidirectional lexical translation probabilities have played a key role in scoring phrase translation rules [20], which are shared by source-to-target and target-to-source translation models. The word alignment model is also improved by agreement using bidirectional alignment information [17]. As the standard NMT directly models the probability of the

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_5

target sentence given a source sentence, it is not trivial to capture bidirectional dependencies.

Several studies have focused on capturing bidirectional dependencies in NMT. Cheng et al. [2] propose bidirectional attention constraints to encourage source-to-target and target-to-source models to agree on word alignment matrices. They achieve the interaction between two directional models through attention mechanisms. Tu et al. [27] train a reconstructor to score target hidden states with the ability to recover the source sentence. This approach is inspired by autoencoders to enhance the adequacy of translations. Monolingual corpora are also successfully incorporated into NMT to augment the limited parallel data through the interaction between two directional translation models [3, 8]. Cheng et al. [3] propose autoencoders to reconstruct monolingual corpora with one translation model as an encoder and the other inverse translation model as a decoder. Meanwhile, [8] leverage reinforcement learning to achieve the interaction between two directional translation models. These two approaches need to translate monolingual corpora first so that they utilize the bidirectional information through discrete sentences. However, estimating the gradients over discrete variables via approaches such as reinforcement learning results in high variances [18], and Monte Carlo sampling process is time-consuming.

In this work, we propose a bidirectional NMT model which connects the source-to-target and target-to-source translation models through a soft approach. The bidirectional NMT models the reconstruction probability of the copy of the source sentence given a source sentence with its corresponding target sentence as a latent bridge. In our approach, we directly and jointly optimize the union parameters of two directional translation models because the whole network architecture is differentiable. Inspired by contrastive learning [7, 19, 24], we also propose an alternative training approach for the bidirectional NMT model which contrastively maximizes the reconstruction probability with its true corresponding target sentence as its latent bridge against those with noise target sentences as latent bridges. The main contributions are summarized as follows:

1. We propose a bidirectional NMT model that opens up the interaction between the source-to-target and target-to-source translation models through a soft connection approach. Transparent to network architecture, it can be applied to arbitrary NMT systems.
2. We propose a contrastive training approach which maximizes the reconstruction probability of true sentence pairs against noise sentence pairs to enhance the information interaction between two directional NMT models.
3. Experiments on Chinese-English and English-German translation tasks show that our approach can achieve significant improvements in both directions over strong NMT and SMT baseline systems.

5.2 Unidirectional Neural Machine Translation

Given a source sentence $\mathbf{x} = \mathbf{x}_1, ...\mathbf{x}_n, ..., \mathbf{x}_N$ and a target sentence $\mathbf{y} = \mathbf{y}_1, ...,$ $\mathbf{y}_m, ..., \mathbf{y}_M$, the start-of-the-art NMT is a neural network based on the encoder-decoder that directly maximizes the conditional probability $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{xy})$ where $\boldsymbol{\theta}_{xy}$ is the parameter set of the neural network. We refer to it as *x-to-y* NMT.

The encoder first reads the source words and converts the discrete words into continuous word embeddings. For example, the embedding $\mathbf{e}_{\mathbf{x}_i}$ of word $\mathbf{x}_i$ is $\mathbf{E}_\mathbf{x}[\mathbf{x}_i]$, where $\mathbf{E}_\mathbf{x}$ is the word embedding table of source words. We denote the v-th column of a matrix as $[v]$. The word embeddings of the source sentence are then encoded into a sequence of hidden states $\mathbf{h} = \mathbf{h}_1, ..., \mathbf{h}_n, ..., \mathbf{h}_N$ using a bidirectional RNN.

The decoder also adopts a RNN to model the conditional probability of the target sentence:

$$P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy}) \propto \exp\{g(\mathbf{y}_{t-1}, \mathbf{s}_t, \mathbf{c}_t; \boldsymbol{\theta}_{xy})\} \tag{5.1}$$

where $\mathbf{s}_t$ is a RNN hidden state at time t, $\mathbf{c}_t$ is a context vector, and $\mathbf{y}_{<t} = \mathbf{y}_1, ..., \mathbf{y}_{t-1}$. The context vector $\mathbf{c}_t$ is weighted sum of every source annotations in $\mathbf{h}$.

It is also straightforward to build a *y-to-x* NMT by treating $\mathbf{y}$ as source and $\mathbf{x}$ as the target. As this type of NMT has particular directional property, we call it unidirectional NMT.

5.3 Bidirectional Neural Machine Translation

The unidirectional NMT generally builds two independent translation models, *x-to-y* and *y-to-x*, where there is no information exchanged between them. Due to the structural divergence between natural languages, the information given by the *y-to-x* translation model can benefit the *x-to-y* translation model. Thus it is appealing to integrate the network architectures and training approaches of translation models in both directions.

We propose a bidirectional NMT model which combines *x-to-y* and *y-to-x* NMT models. A soft connection approach is adopted to open up the direct interaction between two directional models. As shown in Fig. 5.1, the architecture contains two unidirectional NMT. The standard *x-to-y* NMT transforms a source sentence $\mathbf{x}$ to a sequence of embeddings which are computed by:

$$\mathbf{e}_{\hat{\mathbf{y}}_t} = \sum_{i}^{|Y|} \hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy})\mathbf{E}_{\hat{\mathbf{y}}}[y_i] \tag{5.2}$$

where the vocabulary size of the language y is $|Y|$. As $\mathbf{e}_{\hat{\mathbf{y}}_t}$ is weighted sum of the source word embedding matrix $\mathbf{E}_{\hat{y}}$ of the *y-to-x* NMT model, we refer to it as

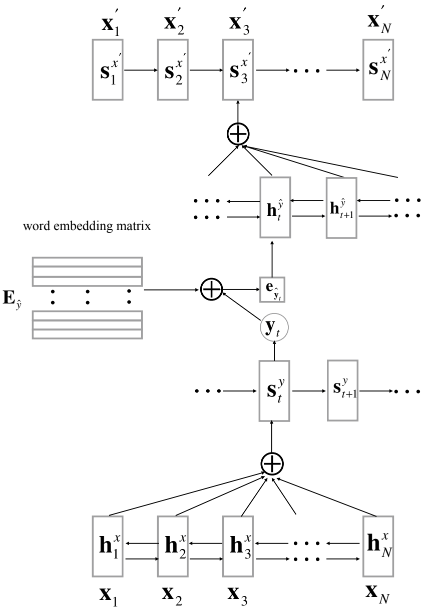


Fig. 5.1 The network architecture for bidirectional NMT. The variable $\mathbf{y}_n$ in the circle node represents the present probability distribution

an expected embedding. $\hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy})$ is the probability of word y_i, which is defined as:

$$\hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy}) \propto \exp\{g(\mathbf{y}_{t-1}, \mathbf{s}_t, \mathbf{c}_t; \boldsymbol{\theta}_{xy})\}^{\alpha} \tag{5.3}$$

Note that we introduce a parameter α to allow us to control the sharpness of the distribution.

$$\mathbf{e}_{\hat{\mathbf{y}}_t} = \sum_{i}^{|Y|} \hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy})\mathbf{E}_{\hat{\mathbf{y}}}[y_i] \tag{5.4}$$

where the vocabulary size of the language y is $|Y|$. As $\mathbf{e}_{\hat{\mathbf{y}}_t}$ is weighted sum of the source word embedding matrix $\mathbf{E}_{\hat{y}}$ of the *y-to-x* NMT model, we refer to it as an expected embedding. $\hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy})$ is the probability of word y_i, which is defined as:

$$\hat{P}(y_i|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}_{xy}) \propto \exp\{g(\mathbf{y}_{t-1}, \mathbf{s}_t, \mathbf{c}_t; \boldsymbol{\theta}_{xy})\}^{\alpha} \tag{5.5}$$

Note that we introduce a parameter α to allow us to control the sharpness of the distribution.

Specifically, given a sentence pair $(\mathbf{x}, \mathbf{y})$, the source sentence $\mathbf{x} = \mathbf{x}_1, .., \mathbf{x}_n, ..., \mathbf{x}_N$ is mapped to an embedding sequence $\mathbf{e}_{\hat{\mathbf{y}}} = \mathbf{e}_{\hat{\mathbf{y}}_1}, ..., \mathbf{e}_{\hat{\mathbf{y}}_m}, ..., \mathbf{e}_{\hat{\mathbf{y}}_M}$ using the *x-to-y* NMT model. The embedding sequence for $\mathbf{y}$, $\mathbf{e}_{\mathbf{y}} = \mathbf{E}_y[\mathbf{y}_1], ..., \mathbf{E}_y[\mathbf{y}_m], ..., \mathbf{E}_y[\mathbf{y}_M]$, is provided for the next step of RNN in the decoder of the *x-to-y* NMT, where $\mathbf{E}_y$ is the target word embedding matrix in the *x-to-y* NMT. The constructed embedding sequence $\mathbf{e}_{\hat{\mathbf{y}}}$ is fed into the *y-to-x* NMT as input. For the *y-to-x* NMT model, it takes $\mathbf{e}_{\hat{\mathbf{y}}}$ as input and $\mathbf{x}'$ as output. $\mathbf{x}'$ is a copy of $\mathbf{x}$. Some similar soft connection approaches are also proposed in other works. Zhang et al. [29] use a similar technique to connect generative and discriminative networks for generating text with adversarial training. Kočiskỳ et al. [12] adopt a reparametrization trick to draw similar $\mathbf{e}_{\hat{\mathbf{y}}}$ from the logistic normal. These soft connection approaches allow us to efficiently backpropagate training gradients.

More formally, the basic idea of bidirectional NMT is inspired by autoencoders. The *x-to-y* NMT is responsible for transforming $\mathbf{x}$ to $\mathbf{e}_{\hat{\mathbf{y}}}$ and the *y-to-x* NMT is used to reconstruct $\mathbf{x}'$. Given a sentence pair $(\mathbf{x}, \mathbf{y})$, the bidirectional NMT model defines a reconstruction function $\mathbf{x} \overset{\mathbf{y}}{\mapsto} \mathbf{x}'$ with the latent variable $\mathbf{y}$ as a bridge if $\mathbf{x}$ is a reconstructed object. As the generation for $\mathbf{e}_{\hat{\mathbf{y}}}$ depends on $\mathbf{x}, \mathbf{y}$ and the parameter set of the *x-to-y* NMT, we formulate the output of this process with a function $\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy})$. Let $R(\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})$ be the output of the bidirectional NMT model which estimates a probability for reconstructing $\mathbf{x}'$ given $\mathbf{x}$ and $\mathbf{y}$. It is computed as the product of reconstruction probabilities for all words in $\mathbf{x}'$.

$$\begin{aligned} & R(\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx}) \\ = & \prod_j P(\mathbf{x}'_j | \mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}); \boldsymbol{\theta}_{yx}) \end{aligned} \tag{5.6}$$

Conditioning on parallel corpora, $\mathbf{y}$ and $\mathbf{x}$ have been already observed . Therefore, we can directly maximize the reconstruction probability $R(\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})$.

In practice, we also jointly incorporate $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{xy})$ and $P(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}_{yx})$ into our training objective to avoid that parameters are trained divergently. It is also natural to combine the reconstruction probability for $\mathbf{y} \overset{\mathbf{x}}{\mapsto} \mathbf{y}'$, $R(\mathcal{O}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{yx}), \mathbf{y}', \boldsymbol{\theta}_{xy})$, if we swap the roles of the *y-to-x* and *x-to-y* NMT. Thus, given a parallel corpus $\mathcal{D} = \{\langle \mathbf{x}^{(d)}, \mathbf{y}^{(d)} \rangle\}_{d=1}^{D}$, we define our training objective as:

$$\begin{aligned} & J(\boldsymbol{\theta}_{xy}, \boldsymbol{\theta}_{yx}) = \\ & \sum_{d=1}^{D} \log P(\mathbf{y}^{(d)} | \mathbf{x}^{(d)}; \boldsymbol{\theta}_{xy}) + \\ & \sum_{d=1}^{D} \log P(\mathbf{x}^{(d)} | \mathbf{y}^{(d)}; \boldsymbol{\theta}_{yx}) + \\ & \lambda_1 \sum_{d=1}^{D} \log R(\mathcal{O}(\mathbf{x}^{(d)}, \mathbf{y}^{(d)}, \boldsymbol{\theta}_{xy}), \mathbf{x}'^{(d)}, \boldsymbol{\theta}_{yx}) + \\ & \lambda_2 \sum_{d=1}^{D} \log R(\mathcal{O}(\mathbf{y}^{(d)}, \mathbf{x}^{(d)}, \boldsymbol{\theta}_{yx}), \mathbf{y}'^{(d)}, \boldsymbol{\theta}_{xy}) \end{aligned} \tag{5.7}$$

where λ_1 and λ_2 are used to balance the importance between the independent probabilities in unidirectional NMT and reconstruction probabilities in bidirectional NMT. We refer to this objective as RECONSTRUCTION LIKELIHOOD.

In bidirectional NMT, $\mathbf{y}$ plays a key role in connecting *x-to-y* and *y-to-x* models. While we can train a fine model to match the training data by maximizing the reconstruction likelihood in Eq. (5.7), it suffers from a problem: an unintended $\tilde{\mathbf{y}}$ may occupy a higher reconstruction probability than the real $\mathbf{y}$. This is partly because Eq. (5.7) may teach the model to focus more on explaining common correlations between $\mathbf{x}$ and observed $\mathbf{y}$ and ignore distinguishing from unobserved $\tilde{\mathbf{y}}$.

We introduce contrastive learning for bidirectional NMT to handle this issue. The intuition is simple: we enhance the reconstruction probability with the ground truth $\mathbf{y}$ as a latent variable against those with noise $\tilde{\mathbf{y}}$s as latent variables.

Inspired by noise contrastive estimation [7] and negative sampling [19], we propose a contrastive learning approach for $\mathbf{x} \overset{\mathbf{y}}{\mapsto} \mathbf{x}'$:

$$\log R(\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx}) + \mathbb{E}_{\tilde{\mathbf{y}} \sim p_n(\mathbf{y})} \left[\log(1 - R(\mathcal{O}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})) \right] \quad (5.8)$$

The basic idea is that we distinguish the real $\mathbf{y}$ from noise $\tilde{\mathbf{y}}$s drawn from a noise distribution $p_n(\mathbf{y})$. As Fig. 5.2 shows, a source sentence "*qiyue jiao zhengchang yu duo.*" and a target sentence "*July was wetter than normal.*" are given. Since the true sentence pair is only observed in training, the biased noise sentences, like "*July was direr than usual.*", may possess a higher reconstruction probability. We want to penalize these noise sentences. The new training objective is formulated by replacing the last two terms of Eq. (5.7) with their corresponding contrastive scores in Eq. (5.8). We refer to the new objective as CONTRASTIVE LIKELIHOOD.

The reconstruction probability $R(\mathcal{O}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})$ for the true sentence pair through bidirectional NMT is calculated at the word level. We therefore also penalize the reconstruction probability for each noise pair at the word level. The last term of Eq. (5.8) is repalced by:

$$\mathbb{E}_{\tilde{\mathbf{y}} \sim p_n(\mathbf{y})} \left[\sum_j \log(1 - R(\mathcal{O}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\theta}_{xy}), \mathbf{x}'_j, \boldsymbol{\theta}_{yx})) \right] \quad (5.9)$$

Lamb et al. [15] propose an adversarial training method, Professor Forcing, to make the generative behaviour and the teacher-forced behaviour closely match. Our training objective Eq. (5.8) is similar to theirs. But if their method is applied to bidirectional NMT, $\boldsymbol{\theta}_{xy}$ are learned aiming to fool the discriminator $\boldsymbol{\theta}_{yx}$ through maximizing $R(\mathcal{O}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})$, while $\boldsymbol{\theta}_{xy}$ and $\boldsymbol{\theta}_{yx}$ in our approach work together to distinguish $\mathbf{y}$ from $\tilde{\mathbf{y}}$ through minimizing $R(\mathcal{O}(\mathbf{x}, \tilde{\mathbf{y}}, \boldsymbol{\theta}_{xy}), \mathbf{x}', \boldsymbol{\theta}_{yx})$.

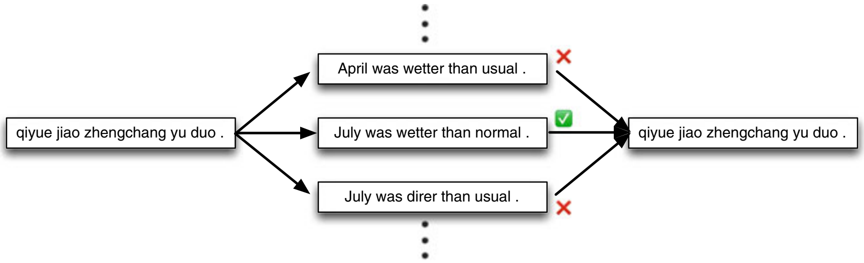

April was wetter than usual .
qiyue jiao zhengchang yu duo .
July was wetter than normal .
July was direr than usual .
✗
✓
✗
qiyue jiao zhengchang yu duo .

Fig. 5.2 The illustration of contrastive learning for the reconstruction probability. The true parallel sentence is (“qiyue jiao zhengchang yu duo.”, “July was wetter than normal.”). We generate negative samples, such as “July was direr than usual”, to contrastively maxmize the reconstruction probability of the source sentence with its true corresponding target sentence as a latent bridge

5.4 Decoding Strategies

After jointly learning the parameter sets $\boldsymbol{\theta}_{xy}$ and $\boldsymbol{\theta}_{yx}$, we adopt two different decoding strategies to obtain the best translation for an input sentence.

1. *Independent decoding*: As the *x-to-y* and *y-to-x* NMT are connected via the reconstruction likelihood and contrastive likelihood, they can hopefully benefit each other and improve their respective parameter learning. Based on the *x-to-y* NMT with its trained parameter $\boldsymbol{\theta}_{xy}$, we apply the standard beam search to find its best translation given an input $\mathbf{x}$ and normalize its cost by candidate length. Similarly, we use the *y-to-x* NMT to decode an input $\mathbf{y}$.
2. *Joint decoding*: The reconstruction score defined by bidirectional NMT contains the information of two directional NMT models which is conducive to reinforcing the confidence of translation results. Therefore, for each candidate translation of an n-best list obtained from independent decoding, we rescore it by combining the prediction score of bidirectional NMT. Further, we can linearly interpolate the four terms in Eq. (5.7) to approximate our training objective. The hyperparameters of linear interpolation can be tuned by MERT [21] in the validation dataset.

5.5 Experiments

5.5.1 Setup

We evaluated our approach on the Chinese-English and English-German translation tasks.

For Chinese-English, the training corpus from LDC consists of 1.25M sentence pairs with 27.9M Chinese words and 34.5M English words respectively. We used the NIST 2006 set as the validation set for hyper-parameter optimization and model selection, while used the NIST 2002, 2003, 2004, 2005, and 2008 as test sets. In the validation and test sets, each Chinese sentence has four English reference sentences. For English-to-Chinese translation, we used them in a reverse direction: the first reference English sentence was treated as a source sentence, and its corresponding Chinese sentence served as the single reference translation. We used the case-insensitive 4-gram BLEU [22] score as the evaluation metric as calculated by `multi-bleu.perl` script.

For English-German, we used WMT 14 training corpus that contains 4M sentence pairs with 91M English words and 87M German words. The size of WMT 14 dataset is a little smaller than WMT 15 dataset. The validation set is newstest2013, and the test sets are newstest2014 and newstest2015. The case-sensitive BLEU score on newstest2015 was calculated using `mteval-v13a.pl` script. The other sets were evaluated by `multi-bleu.perl` script.

We compared our approach with state-of-the-art SMT and NMT systems:

1. MOSES [13]: a phrase-based SMT system;
2. RNNSEARCH [1]: an attention-based NMT system.

We used the default setting to build the phrase-based translation system on the parallel corpus wherein a 4-gram language model was trained by the SRILM toolkit [25]. For RNNSEARCH, we trained the model with the sentences of length up to 50 words. In Chinese-English, the vocabulary size was set to 30 K for both Chinese and English languages, and we also used dropout technique [9]. In English-German, we used the subword units technique [23] to split English and German words into about 50 K subwords. We drew λ and α exponentially from 0.01 to 10. We set λ_1 and λ_2 to 1.0 for reconstruction likelihood and 0.1 for contrastive likelihood. The α was set to 1.0. We find that sampling one noise sample suffices to obtain a good result considering training efficiency and $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{xy})$ is empirically a proper noise distribution. We trained the reconstruction likelihood model from scratch, then initialized the contrastive likelihood model with the parameters of the reconstruction model.

5.5.2 *Effect of Translation Strategies*

We investigate the translation performance of two decoding strategies on different beam sizes. Table 5.1 shows the BLEU scores of joint decoding versus independent decoding.

In independent decoding, increasing beam size does not result in any further improvements on the BLEU scores. But we find both REC. LIKELIHOOD and CON. LIKELIHOOD achieve significant gains over RNNSEARCH (up to +3.0 BLEU), which verifies that the interaction between two directional translation models indeed benefits the improvements of their respective parameters. We also find that CON. LIKELIHOOD performs better than REC. LIKELIHOOD. When adopting joint decoding, our

Table 5.1 Comparison of different decoding strategies and beam size on different models

Model	Indep. (b=10)	Indep. (b=100)	Joint (b=10)	Joint (b=100)
RNNSearch	34.87	34.71	N/A	N/A
Joint Likelihood	35.86	35.81	36.38	36.94
Con. Likelihood	37.81	37.81	38.21	38.60

two approaches achieve new improvements compared with independent decoding. Further improvements can be obtained with a larger beam size. Although joint decoding can lead to extra overhead, we find that this overhead is acceptable. The decoding speed of joint decoding is averagely 0.52 second per sentence compared with 0.38 second per sentence for independent decoding. By default, we adopt joint decoding with beam size 10 for Chinese-English experiments and 12 for English-German experiments.

5.5.3 Comparison with SMT and Standard NMT

Table 5.2 shows the comparison results of our work against SMT and the standard unidirectional NMT. RNNSearch has completely outperformed Moses in both directions. Rec. Likelihood obtain significant improvements on all the datasets of both directions except for NIST04 and NIST05 in English-to-Chinese translation. However, Con. Likelihood can achieve substantial improvements over all the models. In Chinese-to-English translation, Con. Likelihood yields up to 3.5 BLEU gain

Table 5.2 Comparison with Moses, RNNSearch, [27] and [3]. "C→E" denotes Chinese-to-English translation task, "E→C" denotes English-to-Chinese translation task. "NIST06" is the validation set and "NIST02-05" are test sets. The BLEU scores are case-insensitive. "*": significantly better than MOSES ($p < 0.05$); "**": significantly better than MOSES ($p < 0.01$); "+": significantly better than RNNSearch ($p < 0.05$); "++": significantly better than RNNSearch ($p < 0.01$)

Model	Direction	NIST06	NIST02	NIST03	NIST04	NIST05	NIST08
Moses	C → E	31.83	32.61	31.44	32.70	30.13	24.71
	E → C	14.45	18.88	14.53	14.35	13.10	11.23
RNNSearch	C →E	34.87	36.84	34.56	37.56	34.54	27.36
	E →C	18.12	24.39	18.07	18.88	16.96	13.48
Tu et al. [27]	C →E	35.19	37.35	–	–	34.88	27.93
Cheng et al. [3]	C →E	36.16	38.52	36.05	38.38	35.53	28.06
	E →C	18.18	23.87	18.02	18.65	16.56	12.78
Joint Likelihood	C → E	36.38^{**++}	38.51^{**++}	35.91^{**++}	38.89^{**++}	36.62^{**++}	27.79^{**+}
	E →C	19.21^{**++}	24.77 $^{**+}$	19.01^{**++}	18.79**	17.18**	14.12^{**++}
Con. Likelihood	C →E	38.21^{**++}	39.52^{**++}	36.98^{**++}	40.44^{**++}	38.00^{**++}	29.16^{**++}
	E →C	20.01^{**++}	25.67^{**++}	19.91^{**++}	19.42^{**++}	17.88^{**++}	14.89^{**++}

over RNNSEARCH and up to 7.9 BLEU gain over MOSES. In English-to-Chinese translation, significant improvements are also observed. We can find that both of our approaches are capable of using bidirectional information to improve NMT models, while CON. LIKELIHOOD generally performs better than REC. LIKELIHOOD.

We also compare our approaches with [27] and [3]. The experimental results are shown in Table 5.2 wherein the results of Tu et al. [27] are reported from their paper. Because [3] incorporate monolingual corpora into NMT, we implement the autoencoder approach [3] on parallel corpus for comparison fairness. Both our two approaches achieve substantial improvements over [27]. Although REC. LIKELIHOOD achieves comparable results compared with [3] in Chinese-to-English translation, CON. LIKELIHOOD significantly outperforms them in both directions.

5.5.4 BLEU Scores Over Sentence Length

Figure 5.3 shows the change curve of BLEU scores over different source sentence lengths. RNNSEARCH stays nearly consistently better than MOSES for all lengths. However, the gap between them has become smaller when the sentence length is larger than 50 words. One reason for this, we think, is that the length of training sentences is limited to 50 words for efficiency. Our two approaches achieve better translation performance for source sentences with less than 50 words. Unfortunately, the translation performance of CON. LIKELIHOOD drops a little too fast with longer sentences. This may be because the length of negative samples generated for contrastive learning is usually longer than the true data. We leave this problem for future work.

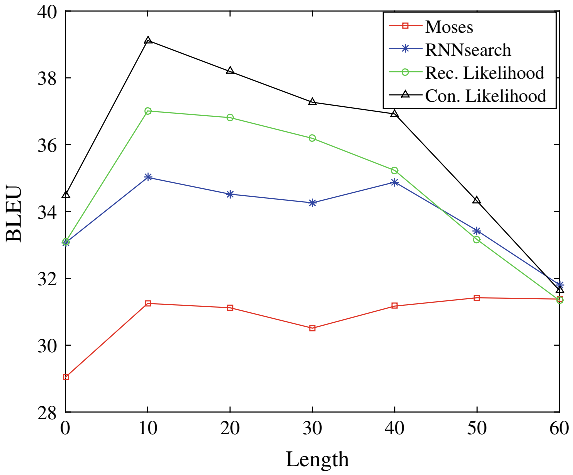


Fig. 5.3 Translation qualities of different systems on different source sentence lengths

Fig. 5.4 Learning curves of different models on the Chinese-to-English validation set

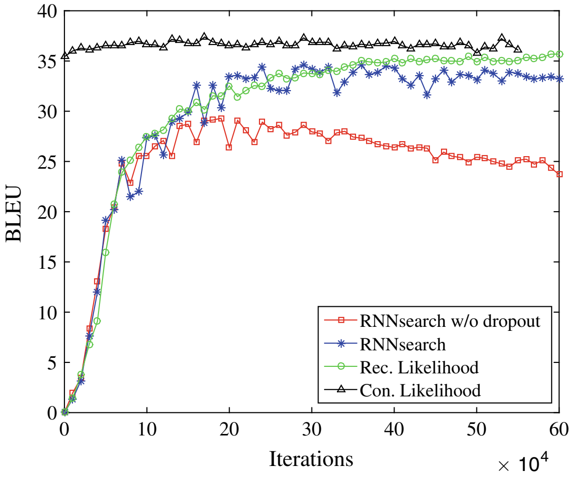


Table 5.3 Example words which are respectively top-five closest to the corresponding target words in the L2 distance

Source: Lipeng huijian lamei jingji tixi changren mishu

Target: Chinese premier meets LAES permanent secretary

Chinese	Premier	Meets	LAES	Permanent	Secretary
Chinese	Premier	Meets	Latin	Permanent	Secretary
China’s	Chancellor	Linages	Mirs	Permanently	Undersecretary
The Chinese	Vice-Premier	Overlap	Beibu	Perpetual	Secretaroes
Chinese	Minster	Attuned	Littoral	Ceremonious	Procurator
Gac	Treasurer	Contracdicts	Ching	Barbarous	Commissioner

5.5.5 Comparison of Learning Curves

We compare the learning curves of four models, RNNSEARCH, REC. LIKELIHOOD, CON. LIKELIHOOD and RNNSEARCH without dropout technique in Fig. 5.4. Without dropout technique [9], RNNSEARCH easily overfits after about 250K iterations. Dropout is able to make the BLEU scores of RNNSEARCH rise steadily and prevent the model from overfitting. It is interesting that we can find that the learning curve of REC. LIKELIHOOD is more stable while it learns slightly slower than RNNSEARCH. And it becomes more robust in terms of BLEU scores with the number of iteration increasing. REC. LIKELIHOOD acts as a regularizer for two directional NMT models to capture bidirectional dependencies. CON. LIKELIHOOD demonstrates its ability to further improve the translation performance. Its learning curve is also very stable.

Table 5.4 Comparison with RNNSEARCH and [10, 11]. "D→E" denotes German-to-English translation task, "E→D" denotes English-to-German translation task. "*": significantly better than RNNSEARCH ($p < 0.05$); "**": significantly better than RNNSEARCH ($p < 0.01$); "+": significantly better than [10, 11] ($p < 0.05$); "++": significantly better than [10, 11] ($p < 0.01$)

Model	Direction	newstest2014	newstest2015
RNNSEARCH	D → E	24.45	27.25
	E →D	19.77	22.63
Jean et al. [10, 11]	D → E	–	25.60
	E →D	19.40	22.40
JOINT LIKELIHOOD	D → E	24.77^{**}	27.54^{*++}
	E →D	19.76^{++}	22.73^{*++}
CON. LIKELIHOOD	D →E	26.15^{**}	28.96^{**++}
	E →D	20.84^{**++}	23.90^{**++}

5.5.6 Analysis of Expected Embeddings

When connecting *x-to-y* and *y-to-x* NMT, we use a soft approach. Equation (5.4) indicates that expected embeddings represent the weighted sum of the language y embedding matrix. In this section, we analyze what they represent. For a given sentence pair $\mathbf{x}$ and $\mathbf{y}$, we can obtain a sequence of expected embeddings $\mathbf{e}_{\hat{\mathbf{y}}}$. We calculate the L2 distance between the expected embedding of each target word in $\mathbf{y}$ and each word embedding in $\mathbf{E}_{\hat{\mathbf{y}}}$. We list the top-five closest words in Table 5.3. It is possible that the expected embedding is capable of retrieving its relevant words. For example, words, like "amid", "amidst" and "amongst", have similar meanings with "during". It seems that the present probability influences the expected embeddings, which results in more diverse expression ability.

5.5.7 Results on English-German Translation

Table 5.4 shows the results on English-German translation. We compare our approaches with both RNNSEARCH and [10, 11]. We observe that our approaches achieve significant improvements in both directions. This suggests that our approaches can be applied to more language pairs. In addition, we find that the contribution from joint decoding is just at most 0.1 BLEU. We conjecture that small beam size, the single reference, and less diverse decoding affect its performance [16].

5.6 Summary

We present a novel bidirectional NMT model. The parameters of these two directional models can be jointly optimized. We also propose a contrastive learning approach for the bidirectional NMT model. The experiments on Chinese-English and German-English translation tasks validate the effectiveness of our approach. It is possible that we can apply our approach to more tasks. In addition, constructing a good noise distribution seems to be an interesting research direction.

References

1. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
2. Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Agreement-based joint training for bidirectional attention-based neural machine translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
3. Cheng, Y., Xu, W., He, Z., He, W, Wu, H., Sun, M., & Liu, Y. (2016) Semi-supervised learning for neural machine translation. In *Association for Computational Linguistics (ACL)*.
4. Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Association for Computational Linguistics (ACL)*.
5. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
6. Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. N. (2016). A convolutional encoder model for neural machine translation. arXiv preprint arXiv:1611.02344.
7. Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*.
8. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T. -Y., & Ma, W. -Y. (2016). Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS)*.
9. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.
10. Jean, S., Cho, K., Memisevic, R., Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Association for Computational Linguistics (ACL)*.
11. Jean, S., Firat, O., Cho, K., Memisevic, R., & Bengio, Y. (2015). Montreal neural machine translation systems for WMT15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.
12. Kočiskỳ, T., Melis, G., Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., Hermann, K. M. (2016). Semantic parsing with semi-supervised sequential autoencoders. In *Empirical Methods in Natural Language Processing (EMNLP)*.
13. Koehn, P., & Hoang, H. (2007). Factored translation models. In *Empirical Methods in Natural Language Processing (EMNLP)*.
14. Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *North American Association for Computational Linguistics (NAACL)*.
15. Lamb, A. M., Goyal, A. G. A. P., Zhang, Y., Zhang, S., Courville, A. C., & Bengio, Y. (2016). Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems (NIPS)*.

16. Li, J., Monroe, W., & Jurafsky, D. (2016). A simple, fast diverse decoding algorithm for neural generation. arXiv preprint arXiv:1611.08562.
17. Liang, P., Taskar, B., & Klein, D. (2006). Alignment by agreement. In *North American Association for Computational Linguistics (NAACL).*
18. Maddison, C. J., Mnih, A., Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712.
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS).*
20. Och, F. (2003). Minimum error rate training in statistical machine translation. In *Association for Computational Linguistics (ACL).*
21. Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational linguistics*.
22. Papineni, K., Roukos, S., Ward, T., Zhu, W. -J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL).*
23. Sennrich, R., Haddow, B., Birch, A. (2016). Neural machine translation of rare words with subword units. In *Association for Computational Linguistics (ACL).*
24. Smith, N. A., Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Association for Computational Linguistics (ACL).*
25. Stolcke, A. (2002). SRILM—An extensible language modeling toolkit. In *International Conference on Spoken Language Processing*.
26. Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS).*
27. Tu, Z., Liu, Y., Shang, L., Liu, X., & Li, H. (2017). Neural machine translation with reconstruction. In *Association for the Advancement of Artificial Intelligence (AAAI).*
28. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., & Macherey, K. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.
29. Zhang, Y., Gan, Z., & Carin, L. (2016). Generating text via adversarial training.

Chapter 6
Related Work

Abstract In this chapter, we first review attention-based neural machine translation (NMT). Attentional mechanism make NMT outperform the non-attentional models. The related works confirm the importance of attentional mechanisms. Then we investigate how the bidirectional information is captured in SMT and NMT. Next we summarize a number of work which incorporate additional data resources, such as monolingual corpora and pivot language corpora, into machine translation systems. Finally, we make a simple review of the studies about contrastive learning, which is a key technique in our fourth work.

6.1 Attentional Mechanisms in Neural Machine Translation

Bahdanau et al. [2] first introduce the attentional mechanism into neural machine translation to enable the decoder to focus on relevant parts of the source sentence during decoding. The attention mechanism allows a neural model to cope better with long sentences because it does not need to encode all the information of a source sentence into a fixed-length vector regardless of its length. In addition, the attentional mechanism allows us to look into the "black box" to gain insights on how NMT works from a linguistic perspective.

Luong et al. [24] propose two simple and effective attentional mechanisms for neural machine translation and compare various alignment functions. They show that attention-based NMT are superior to non-attentional models in translating names and long sentences.

After analyzing the alignment matrices generated by RNNSEARCH [2], we find that modeling the structural divergence of natural languages is so challenging that unidirectional models can only capture part of alignment regularities. This finding inspires us to improve attention-based NMT by combining two unidirectional models. In the first work, we only apply agreement-based joint learning to RNNSEARCH. As our approach does not assume specific network architectures, it is possible to apply it to the models proposed by Luong et al. [24].

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_6

6.2 Capturing Bidirectional Dependencies

6.2.1 Capturing Bidirectional Dependencies

Capturing bidirectional dependencies has shown promising results to benefit the independent models. In SMT, people have developed a number of approaches, like bidirectional lexical probabilities [26], word alignment by agreement [21], bilingual language model [12] and so on, to improve the performance of independent SMT. In NMT, several authors have also tried to capture bidirectional dependencies [5, 6, 15, 19, 31]. Similarly, [18] apply semi-supervised sequential autoencoders to incorporating unsupervised training data in semantic parsing.

Closely related to [6, 15], our approach in the fourth work connects two directional models within the parameter space and also jointly trains them. In addition, our approach aims at capturing bidirectional dependencies on parallel corpora while theirs try to exploit monolingual corpora. In contrast to [31], our approach does not require a new-designed reconstructor module. We also propose a contrastive learning approach to enhance bidirectional dependencies.

6.2.2 Agreement-Based Learning

As one reprensentative work for capturing bidirectional information, [21] first introduce agreement-based learning into word alignment: encouraging asymmetric IBM models to agree on word alignment, which is a latent structure in word-based translation models [4]. This strategy significantly improves alignment quality across many languages. They extend this idea to deal with more latent-variable models in grammar induction and predicting missing nucleotides in DNA sequences [22].

Liu et al. [23] propose generalized agreement for word alignment. The new general framework allows for arbitrary loss functions that measure the disagreement between asymmetric alignments. The loss functions can not only be defined between asymmetric alignments but also between alignments and other latent structures such as phrase segmentations.

In attention-based NMT, word alignment is treated as a parametrized function instead of a latent variable. This makes word alignment differentiable, which is important for training attention-based NMT models. Although alignment matrices in attention-based NMT are in principle “symmetric” as they allow for many-to-many soft alignments, we find that unidirectional modeling can only capture partial aspects of structure mapping. The contribution of our first work is to adapt agreement-based learning into attentional NMT, which significantly improves both alignment and translation.

6.3 Incorporating Additional Data Resources

6.3.1 Exploiting Monolingual Corpora for Machine Translation

Exploiting monolingual corpora for conventional SMT has attracted intensive attention in recent years. Several authors have introduced transductive learning to make full use of monolingual corpora [3, 32]. They use an existing translation model to translate unseen source text, which can be paired with its translations to form a pseudo parallel corpus. This process iterates until convergence. While [17] propose an approach to estimating phrase translation probabilities from monolingual corpora, [38] directly extract parallel phrases from monolingual corpora using retrieval techniques. Another important line of research is to treat translation on monolingual corpora as a decipherment problem [9, 27].

Due to the limit in quantity, quality and coverage for parallel corpora, additional data resources have raised attention in NMT recently. Gulccehre et al. [13] propose to incorporate target-side monolingual corpora as a language model for NMT. Sennrich et al. [28] pair the target monolingual corpora with its corresponding translations, then merge them with parallel corpora for retraining source-to-target model. Zhang and Zong [39] propose two approaches, self-training algorithm and multi-task learning framework, to incorporate source-side monolingual corpora. He et al. [15] also introduce a dual learning algorithm to incorporate monolingual sentences based on reinforcement learning. The proposed model can exploit both source and target monolingual corpora.

6.3.2 Autoencoders in Unsupervised and Semi-supervised Learning

Autoencoders and their variants have been widely used in unsupervised deep learning ([1, 30, 35], just to name a few). Among them, Socher et al. [30]'s approach bears close resemblance to our approach as they introduce semi-supervised recursive autoencoders for sentiment analysis. The difference is that we are interested in making a better use of parallel and monolingual corpora while they concentrate on injecting partial supervision to conventional unsupervised autoencoders. Dai and Le [8] introduce a sequence autoencoder to reconstruct an observed sequence via RNNs. Our approach differs from sequence autoencoders in that we use bidirectional translation models as encoders and decoders to enable them to interact within the autoencoders.

6.3.3 Machine Translation with Pivot Languages

Machine translation suffers from the scarcity of parallel corpora. For low-resource language pairs, a pivot language is introduced to “bridge” source and target languages in statical machine translation [7, 10, 33, 36, 37]. One of the most representative approaches is triangulation approach that a source-to-target phrase table is generated by combining source-to-pivot and pivot-to-target phrase tables.

In NMT, [11, 16] propose multi-way, multilingual NMT models that enable zero-resource machine translation. The multi-way, multilingual NMT model is able to translate one of N languages to one of M languages. where attention mechanism is shared between source-target language pairs. They propose multi-source translation strategy that combines source and pivot language sentences. They also propose fine-tuning with a pesudo source-to-target parallel corpus to directly train a model for source-to-target language pair from multi-way, multilingual model. Zoph et al. [40] adopt transfer learning to fine-tune parameters of the low-resource language pairs using trained parameters on the high-resource language pairs. However, our approach in the third work aims to jointly train source-to-pivot and pivot-to-target NMT models, which can alleviate the error propagation of pivot-based approaches. We use connection terms to “bridge” these two models and make them benefit each other.

6.4 Contrastive Learning

Contrastive learning is a popular technique in a variety of fields. Gutmann and Hyvärinen [14] propose noise contrastive estimation (NCE) to estimate unnormalized statistical models where we can not accurately compute the partition function. They introduce noise examples to be discriminated from observed examples. A simplified extension, negative sampling (NEG), is developed for learning distributed vector representations [25]. There are also many other works that focus on avoiding computing the partition function with contrastive objectives [20, 29, 34].

References

1. Ammar, W., Dyer, C., & Smith, N. (2014). Conditional random field autoencoders for unsupervised structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*.
2. Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
3. Bertoldi, N., & Federico, M. (2009). Domain adaptation for statistical machine translation. In *Workshop On Statistical Machine Translation*.
4. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.

5. Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M, & Liu, Y. (2016). Agreement-based joint training for bidirectional attention-based neural machine translation. In *International Joint Conference on Artificial Intelligence (IJCAI).*
6. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2016). Semi-supervised learning for neural machine translation. In *Association for Computational Linguistics (ACL).*
7. Cohn, T., & Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Association for Computational Linguistics (ACL).*
8. Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems (NIPS).*
9. Dou, Q., Vaswani, A., & Knight, K. (2014). Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Empirical Methods in Natural Language Processing (EMNLP).*
10. El Kholy, A., Habash, N., Leusch, G., Matusov, E., & Sawaf, H. (2013). Language independent connectivity strength features for phrase pivot statistical machine translation. In *Association for Computational Linguistics (ACL).*
11. Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman Vural, F. T., & Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP).*
12. Garmash, E., & Monz, C. (2014). Dependency-based bilingual language models for reordering in statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP).*
13. Gulccehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. arXiv:1503.03535 [cs.CL].
14. Gutmann, M. U., & Hyvärinen, A. (2012) Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*.
15. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T. -Y., & Ma, W. -Y. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS).*
16. Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using lstm for region embeddings. arXiv preprint arXiv:1602.02373.
17. Klementiev, A., Irvine, A., Callison-Burch, C., & Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *European Association for Computational Linguistics (EACL).*
18. Kočiskỳ, T., Melis, G., Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., Hermann, K. M. (2016). Semantic parsing with semi-supervised sequential autoencoders. In *Empirical Methods in Natural Language Processing (EMNLP).*
19. Li, J., Monroe, W., Jurafsky, D. (2016). A simple, fast diverse decoding algorithm for neural generation. arXiv preprint arXiv:1611.08562.
20. Liang, P., & Jordan, M. I. (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning (ICML).*
21. Liang, P., Taskar, B., & Klein, D. (2006). Alignment by agreement. In *North American Association for Computational Linguistics (NAACL).*
22. Liang, P., Klein, D., & Jordan, M. I. (2007). Agreement-based learning. In *Advances in Neural Information Processing Systems (NIPS).*
23. Liu, C., Liu, Y., Luan, H., Sun, M., & Yu, H. (2015). Generalized agreement for bidirectional word alignment. In *Empirical Methods in Natural Language Processing (EMNLP).*
24. Luong, M. -T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP).*
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS).*
26. Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational linguistics*.

27. Ravi, S., & Knight, K. (2011). Deciphering foreign language. In *Association for Computational Linguistics (ACL)*.
28. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving nerual machine translation models with monolingual data. In *Association for Computational Linguistics (ACL)*.
29. Smith, N. A. & Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Association for Computational Linguistics (ACL)*.
30. Socher, R., Pennington, J., Huang, E., Ng, A., & Manning, C. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Empirical Methods in Natural Language Processing (EMNLP)*.
31. Tu, Z., Liu, Y., Shang, L., Liu, X., & Li, H. (2017). Neural machine translation with reconstruction. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
32. Ueffing, N., Haffari, G., & Sarkar, A. (2007). Trasductive learning for statistical machine translation. In *Association for Computational Linguistics (ACL)*.
33. Utiyama, M., & Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *North American Association for Computational Linguistics (NAACL)*.
34. Vickrey, D., Lin, C. C., Koller, D. (2010). Non-local contrastive objectives. In *International Conference on Machine Learning (ICML)*.
35. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. -A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*.
36. Hua, W., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*.
37. Zahabi, S. T., Bakhshaei, S., & Khadivi, S. (2013). Using context vectors in improving a machine translation system with bridge language. In *Association for Computational Linguistics (ACL)*.
38. Zhang, J., & Zong, C. (2013). Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Association for Computational Linguistics (ACL)*.
39. Zhang, J., & Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
40. Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Chapter 7
Conclusion

7.1 Conclusion

Machine translation investigates the use of software to automatically translate text or speech from one language to another language, with no human involved. Research works on machine translation have been widely studied since its first appearance in 1949. The invention of NMT makes related research studies about machine translation come to a new climax. The basic encoder-decoder framework of NMT has also been applied in other research fields, such as image caption [11] and sentence summarization [9]. Because of the substantial improvement of translation performances on human evaluation, many novel techniques about NMT have been successfully adopted in the industrial machine translation service [10].

In this book, we mainly work on four topics in NMT. All our works have one thing in common where two NMT models are involved. We jointly model and train two NMT models to enable them to benefit each other.

The first work connects the source-to-target and target-to-source NMT models through attentional mechanisms. We have presented agreement-based joint training to encourage these two NMT models to agree on parameterised alignment matrices. Because attentional mechanisms are capable of guiding the decoder to find the relevant source words to replenish the context information, we aim to improve the translation quality driven by increasing the accuracy of the attentional mechanism. The experimental results verify that our approach can improve not only the accuracy of the attentional mechanism but also the translation performance.

In the second work, we associate the source-to-target and target-to-source NMT models with an autoencoder. The autoencoder is used to reconstruct monolingual sentences, in which the source-to-target and target-to-source NMT models collaborate to maximize the reconstruction probability. It enables NMT to incorporate monolingual corpora into the standard NMT model. Based on parallel and monolingual corpora, we propose semi-supervised learning for NMT. Experiments on Chinese-English NIST datasets show that our approach leads to significant improvements.

Y. Cheng, *Joint Training for Neural Machine Translation*, Springer Theses,
https://doi.org/10.1007/978-981-32-9748-7_7

In the third work, we shift our focus on zero-resource NMT, which is a more severe challenge. Due to the unavailability of parallel corpora, we introduce a third language to bridge the source-to-target translation. However, the source-to-pivot and pivot-to-target translation models are usually independently trained. We propose three connection terms for a joint model which contains the source-to-pivot and pivot-to-target NMT models. Experiments on German-French and Spanish-French translation tasks with English as the pivot language show that our joint training approach improves the translation quality significantly than independent training on source- to-pivot, pivot-to-target and source-to-target directions.

The fourth work aims to make both the modeling and training of NMT capture bidirectional dependencies. We propose an end-to-end bidirectional NMT model, which involves the source-to-target and target-to-source NMT models. A soft connection approach is adopted to achieve the end-to-end modeling and training, which opens up the interaction between these two directional models. For enhancing the bidirectional information to be shared, we propose an improved training approach, contrastive learning. We contrastively maximize the probability of true data against the noisy data. Experiments on Chinese-English and German-English translation tasks show that our approach can achieve significant improvements in both directions over both the state-of-the-art NMT and SMT systems.

We validate the effectiveness of joint training for NMT by means of four hot topics in NMT. We believe the proper interaction of multiple NMT models can facilitate the information transfer and remedy drawbacks induced by unidirectional models.

7.2 Future Directions

Due to the structural divergence between different language pairs, unidirectional models might only possess the partial information. Joint training for multiple NMT models with different translation directions is beneficial to deliver its own information to each other. Our future work includes three directions.

7.2.1 Joint Modeling

How to develop a good interaction way to open up the information flow among multiple models deserves to be further studied. Our book mainly includes two interaction ways: (1) through discrete variables [2, 4]; (2) through continuous variables [3].

In both the second and the third works, a search space is required to deal with the latent variable problem. Due to the intractability to accurately enumerate all the latent variables, we use a n-best decoding list to approximate this search space. It is clear that we can observe and analyze what these latent variables look like. We can feel free to incorporate some prior knowledge to restrict the search space. However, estimating the gradients through discrete variables using approaches, like reinforcement learning, results in high variances. Generating a sample list incurs massive computational time. Besides, the approximation would become inferior if the size of samples is less. A soft approach which connects multiple models in the parameterised space is a good alternative to circumvent these disadvantages. The first work associates two NMT models on the parameterized attentional mechanisms, which appends a restricted term to guide the alignment matrices to agree on some confident cells. In the fourth work, we propose a soft connection approach which conveys the output information of the source-to-target model to the target-to-source model. The soft connection makes it effective to propagate the error information of the top model back to the bottom model. Zhang et al. [12] and Kočiskỳ et al. [6] have used some similar soft connection techniques to connect multiple models. However, an effective soft connection approach is usually defined for a particular problem. We can not also give some specific analysis for these intermediate outputs.

In the future, we will focus on proposing some more effective approaches that can connect multiple modes more closely.

7.2.2 Joint Training

A good training criterion determines how much performance the model can achieve. Designing some novel training criteria based on the interaction of multiple models is also an interesting research direction. The first work jointly trains two NMT models to encourage the agreement between alignment matrices. The other works are more or less inspired the idea of autoencoders which reconstruct a translation pathway through latent variables.

In recent years, an increasing number of research works on generative adversarial networks have sprung up [5, 7, 8]. A generative adversarial network involves two neural networks, generative and discriminative networks. The generative network captures the data distribution, and the discriminative network is in charge of discriminating a sample came from the training data rather than the generative network. The generative network is used to generate fake samples to fool the discriminator. These two models interact in an adversarial way.

Inspired by this adversarial training, we will apply it to training multiple NMT models. The source-to-target NMT model is treated as a generative model. We aim to train it to better capture the data distribution $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$. The source-to-target model is used to generate some noisy examples $\mathbf{y}'$. With true samples $\mathbf{y}$ and noisy samples $\mathbf{y}'$ given as inputs, we can build a classifier which is responsible for deciding whether they are generated from the source-to-target model or belongs to the set of true parallel data. Usually, a binary classifier that can be a CNN or RNN is a common option for the discriminative network. We will plan to use a target-to-source model as our discriminative network instead. We believe that the target-to-source model expresses more robust to judge whether they are true data or noisy data.

7.2.3 More Tasks

Recently, joint training for multiple models has received much interest on a variety of NLP tasks. Cheng et al. [1] apply agreement-based joint training to the dependency parsing. Kočiskỳ et al. [6] propose a similar autoencoder to incorporate unlabeled data on semantic parsing. We believe that joint training has the potential for more NLP tasks, such as dialogue. In the dialogue system, there exist multiple conversational agents. We can extend the joint training approaches in our works to the dialogue learning.

References

1. Cheng, H., Fang, H., He, X., Gao, J., & Deng, L. (2016). Bi-directional attention with agreement for dependency parsing. In *Empirical methods in natural language processing (EMNLP)*.
2. Cheng, Y., Liu, Y., Yang, Q., Sun, M., & Xu, W. (2016). Neural machine translation with pivot languages. arXiv:1611.04928.
3. Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M., et al. (2016). Agreement-based joint training for bidirectional attention-based neural machine translation. In *International joint conference on artificial intelligence (IJCAI)*.
4. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., et al. (2016). Semi-supervised learning for neural machine translation. In *Association for computational linguistics (ACL)*.
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*.
6. Kočiskỳ, T., Melis, G., Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., et al. (2016). Semantic parsing with semi-supervised sequential autoencoders. In *Empirical methods in natural language processing (EMNLP)*.
7. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv:1511.05644.
8. Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434.
9. Rush, A. M., Chopra, S, & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Empirical methods in natural language processing (EMNLP)*.
10. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.
11. Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning (ICML)*.
12. Zhang, Y., Gan, Z., & Carin, L. (2016). Generating text via adversarial training.