中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

基于模型正则化的神经机器翻译研究

作者姓名：郭登级

指导教师：冯洋 研究员

中国科学院计算技术研究所

学位类别：工学硕士

学科专业：计算机应用技术

培养单位：中国科学院计算技术研究所

2022 年 6 月

Model Regularization for Neural Machine Translation

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Science in Engineering

in Computer Application Technology

By

Dengji Guo

Supervisor: Professor Yang Feng

Institute of Computing Technology, Chinese Academy of Sciences

June, 2014

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

机器翻译是指利用计算机将一种语言自动翻译为另一种语言的技术，是自然语言处理领域最重要的研究方向之一。近 10 年来，随着深度学习的快速发展与广泛应用，神经机器翻译取得了巨大的成功，取代统计机器翻译成为学术研究和业界应用的主流技术。不同于统计机器翻译建立统计数学模型的做法，神经机器翻译基于神经网络建立翻译模型并使用梯度下降算法训练模型。神经机器翻译模型一般包含大量的参数，具有强大的学习能力，但也有容易过拟合的问题。模型往往过度地匹配特定的训练集数据，以至于牺牲了在其他数据上的表现。

为了缓解神经机器翻译模型的过拟合问题，研究者们提出了一系列正则化方法，包括约束模型词向量参数的词向量正则化方法、提高模型噪声鲁棒性的扰动正则化方法和优化模型训练标签的标签平滑方法等。这些方法取得了显著的效果，但也分别存在一些需要改进的问题。为了进一步解决神经机器翻译的过拟合问题，本文针对词向量正则化、扰动正则化和标签平滑三个研究内容分别提出了以下三个方法：

1. **基于语义相关性的词向量正则化方法**

神经机器翻译模型的词向量参数占总参数量的一半左右，具有较大的冗余度。一些正则化方法提出通过共享词表、增强源端-目标端词向量的对齐性等方式来规范词向量参数的学习。然而，这些方法通常依赖相同字符系统下的词形匹配或者外部对齐工具的辅助，在应用范围和使用的灵活性方面有较大的局限性。本文提出一种增强词向量语义相关性的正则化方法，该方法完全基于双语语料的内在特征，适用于所有翻译方向且不依赖外部的知识指导。具体而言，机器翻译的一个训练样本由一个源端句子和一个目标端句子组成，其中隐含了单语言的词共现信息和双语言之间的词共现信息。基于这两种共现信息，本文提出一个自编码的训练目标来同时促进词向量的单语相关性和对齐相关性。实验结果表明，本文的方法极大地增强了词向量参数的语义特征，显著提升了神经机器翻译模型的性能。

2. **基于预测差异的扰动正则化方法**

神经机器翻译模型过拟合的表现之一是对噪声的鲁棒性很差，在输入中添

加少量的噪声就会导致模型的性能出现大幅度的下降。扰动正则化方法通过对训练样本添加噪声以提升模型在噪声数据上的表现，其隐含的前提是模型对原始样本的拟合程度总是高于对被扰动样本的拟合程度，但本文发现该前提并不总是成立。本文利用模型在受到扰动前后对目标词的预测差异分析了词级别样本的拟合情况，发现模型对相当一部分样本是相对欠拟合的，对这部分样本进行扰动正则化训练会损害模型的性能。为了同时缓解模型的过拟合和欠拟合问题，本文提出将输入层扰动引起的预测差异作为正则项训练神经机器翻译模型。在常用的数据集上，本文的方法相比现有的方法取得了巨大的提升，显著地改善了模型对噪声的鲁棒性。

3. **基于条件变分自编码器的标签平滑方法**

双语语料的相对稀缺和使用硬标签训练的方式导致了神经机器翻译模型对已知训练标签过度自信的问题。常规的标签平滑方法使用平均分布对硬标签进行平滑，其先验假设显然不是最优的。对于机器翻译任务而言，每个目标端词的平滑标签应符合其所在的语境。本文利用条件变分自编码器能够基于给定条件将数据标签编码到一个隐变量空间并重构出来的特点，为神经机器翻译设计了一个基于条件变分自编码器的平滑标签生成器。该生成器能够学习给定源端和目标端输入条件下的词级别标签的隐变量分布，并通过该分布实时采样隐变量并生成新标签，用于神经机器翻译模型的在线标签平滑。实验结果表明，本文的方法能够较好地缓解模型对训练标签过度自信的问题并提升模型的翻译性能。

关键词：正则化，词向量，鲁棒性，标签平滑，神经机器翻译

Abstract

Machine translation refers to the technology of using computers to automatically translate one language into another, and it is one of the most important research directions in the field of natural language processing. In the past 10 years, with the rapid development and wide application of deep learning, neural machine translation has achieved great success, replacing statistical machine translation as the mainstream technology for academic research and industry applications. Unlike statistical machine translation, which builds statistical mathematical models, neural machine translation builds models based on neural networks and uses gradient descent to update parameters. Neural machine translation models generally contain a large number of parameters and have powerful fitting capabilities, but they also have the problem of being prone to over-fitting. NMT models tend to over-fit specific samples in the training set, even at the expense of performance on unseen samples.

In order to alleviate the over-fitting problem of neural machine translation models, researchers have proposed a series of regularization methods, including word embedding regularization methods that constrain the word embedding parameters of the model, perturbation regularization methods to improve the models' robustness to noise, and label smoothing methods to optimize the training labels for models, etc. These methods have achieved remarkable performance, but there are also problems that need to be improved. In order to further solve the over-fitting problem of neural machine translation, this paper proposes the following three methods respectively for word vector regularization, perturbation regularization and label smoothing:

1. **Word Embedding Regularization Based on Semantic Relevance**

The word embedding parameters of neural machine translation models account for about half of the total parameters, and have a large redundancy. Some regularization methods propose to regularize the learning of word embedding parameters by sharing vocabulary or enhancing the alignment of source and target word embeddings. However, these methods usually rely on the morphological matching between two languages

sharing the same character system or the assistance of external alignment tools, and have great limitations in terms of application scenario and flexibility. This paper proposes a regularization method to enhance the semantic features of word embeddings, which is completely based on the intrinsic characteristics of bilingual corpora, and is applicable for all translation directions without relying on the guidance of external knowledge. Specifically, a training sample for machine translation consists of a source-side sentence and a target-side sentence, which implies both the monolingual and bilingual word co-occurrence information. Based on these two kinds of co-occurrence information, this paper proposes an auto-encoding training objective to simultaneously promote the monolingual and aligned relevances of word embeddings. Experimental results show that our method can greatly enhance the semantic features of word embedding parameters and significantly improve the performance of neural machine translation models.

2. **Perturbation Regularization Based on Prediction Difference**

One of the manifestations of over-fitting of neural machine translation models is their poor robustness to noise. Adding a small amount of noise to the input will lead to a large drop in the performance of the model. The perturbation regularization methods improves the performance of the model on noisy data by applying input perturbation to the training samples. Its hidden premise is that the a model is always more fitted to the original samples than input-perturbed samples. This paper find that this premise does not always hold. This paper proposes to analyze the fitting of word-level samples by using the prediction difference of the target word before and after the perturbation of the input. It is found that the model is relatively under-fitted to a considerable part of the original samples, and perturbation regularization training for this part of the samples will damage the model's performance. In order to alleviate both the over-fitting and under-fitting problems, this paper proposes to train the neural machine translation model with the prediction difference caused by the input perturbation as an additional regularization term. On widely used datasets, our method achieves huge improvements over existing methods and significantly improves the model's robustness to noise.

3. **Label Smoothing Based on Conditional Variational AutoEncoder**

The relative sparsity of bilingual corpora and the training of hard labels lead to the

overconfidence of neural machine translation models on known training labels. Conventional label smoothing method uses an average distribution to smooth hard labels, but its prior is obviously not optimal. For machine translation tasks, the smoothing label for each target word should match its context. This paper designs a smoothing label generator based on conditional variational autoencoder for neural machine translation by utilizing the characteristics of conditional variational autoencoder that can encode data labels into a latent variable space and reconstruct them based on given conditions. The generator is able to learn a latent variable distribution of word-level labels given source and target inputs, sample latent variables and generate new labels in real-time from this distribution for online label smoothing of neural machine translation models. The experimental results show that our method can better alleviate the overconfidence of training labels and improve the translation performance of the model.

Keywords: Regularization, Word embedding, Robustness, Label smoothing, Neural machine translation

目 录

图形列表

表格列表

第 1 章　引言

1.1　研究背景与意义

机器翻译（Machine Translation，简称 MT）是指利用计算机将一种语言（源语言）自动翻译成另一种语言（目标语言）的技术，是自然语言处理最重要的研究方向之一。在机器翻译出现以前，不同语言背景的人群主要通过掌握多种语言的翻译专家进行交流。第二次世界大战前后，随着国际政治、经济、文化、军事交流的不断增加，人类社会对翻译的需求不断增大，传统上通过人工翻译进行国际交流的方式越来越难以满足现实需求。在此背景下，英国工程师布斯（Booth）和美国工程师韦弗（Weaver）于 20 世纪 40 年代首次提出了利用计算机进行翻译的想法，开启了机器翻译的研究热潮。

机器翻译具有巨大的实用价值。不同于传统人工翻译效率低、成本高的特点，机器翻译利用计算机进行翻译，具有效率高、成本低的先天优势。得益于近年来互联网和移动设备的迅速发展，机器翻译的使用还变得更加灵活便携。现如今，通过“终端-服务器”的“云翻译”模式，人们在个人电脑、手机、翻译机、网页、聊天软件、视频播放器等终端信息载体上的翻译请求能够被及时传递给服务器进行翻译，从而实现高质量的即时翻译。目前大型的互联网公司例如谷歌、微软、腾讯、百度等都已推出在线、免费、覆盖常用语言对的机器翻译系统，一些专注于机器翻译的公司例如科大讯飞、有道翻译等还推出了可以离线使用的便携式翻译机，甚至一些聊天软件和视频平台例如微信、Youtube 等也相继推出了翻译服务，使得聊天记录里的语音和文本、视频中的音频和字幕可以得到实时的翻译。机器翻译的广泛应用，满足了人类日常生活中巨大的翻译需求，体现了机器翻译的实用价值。

机器翻译还具有重要的学术价值。机器翻译技术融合了语言学、计算机学、数学、统计学等诸多学科，是人工智能和自然语言处理最重要的研究方向之一。机器翻译技术的进步，可以促进自然语言处理其它研究方向的发展，甚至能够为机器学习的其它领域的发展提供借鉴和启发。例如最初被用于机器翻译的 Transformer 模型 (Vaswani 等, 2017)，不仅在模型预训练、机器对话等自然语言处理研究领域得到进一步的应用 (Olabiyi 等, 2019; Vlasov 等, 2019)，而且在计算机视觉领域也

产生了重要的影响 (Carion 等, 2020; Dosovitskiy 等, 2021)。

自 20 世纪 50 年代首个机器翻译系统诞生以来，科学家们对机器翻译技术进行了不断的改进。按照建立机器翻译模型的基本原理，机器翻译的发展大概可以分为三个阶段。在 20 世纪 50 年代至 90 年代的发展早期，机器翻译主要基于语言专家总结的规则建立模型（规则机器翻译模型，Rule-Based Machine Translation）。规则机器翻译模型不需要训练，可解释性强，但是严重依赖人为总结的语言规则，维护成本很高，翻译效果较差且难以继续提升。20 世纪 90 年代初，IBM 的 Peter Brown 等人提出了统计机器翻译模型（Statistical Machine Translaiton，SMT）(Brown 等, 1990, 1993)，该模型利用统计数学方法对翻译任务进行建模，能够通过期望最大化算法（Expectation Maximization，EM）自动从大量数据中学习到统计知识。统计机器翻译不依赖人力资源，可解释性强，翻译质量也更加优秀，因此取代规则机器翻译成为该时期的主流方法。21 世纪早期，统计机器翻译从基于词的翻译过渡到基于短语的翻译 (Koehn 等, 2003, 2007a)，并被广泛应用于工业界的在线翻译系统。各大互联网公司研发的早期的机器翻译工具（2014 年以前）都是基于统计机器翻译技术建立的。

进入 21 世纪，随着并行计算硬件设备的进步，深度学习开始被广泛应用于机器翻译，机器翻译进入了神经网络时代。2013 年，Kalchbrenner 和 Blunsom 利用神经网络实现了首个基于 “编码器-解码器” 的架构（Encoder-Decoder Framework）的神经机器翻译模型 (Kalchbrenner 等, 2013)。随后 Sutskever (Sutskever 等, 2014)、Bahdanau (Bahdanau 等, 2015) 等人分别引入 Long Short-Term Memory（LSTM）和注意力机制（Attention），解决了模型的梯度更新和长距离依赖的问题。2017 年，Google 提出了完全基于注意力机制的 Transformer (Vaswani 等, 2017)，不仅加快了神经机器翻译模型训练速度，还大幅提升了翻译性能。自此，Transformer 成为机器翻译最先进、最主流的模型，随后的 4 年里，学术界和工业界对机器翻译的研究和应用基本都是基于 Transformer 模型开展的。

神经机器翻译的优异性能使得机器翻译真正达到了可用的程度，这不仅促进了机器翻译在日常生活中的广泛应用，还使得机器翻译的研究步入了一个新的阶段。一些在过去因为难度过大而不受关注的研究方向逐渐成为当前的研究热点，例如多语言机器翻译 (Dong 等, 2015; Luong 等, 2016; Firat 等, 2016; Lee 等, 2017)、同声传译 (Cho 等, 2016; Gu 等, 2017; Ma 等, 2019) 和多模态机器翻译

(Calixto 等, 2017; Zhao 等, 2020; Madhyastha 等, 2017) 等。同时作为一项新的技术，神经机器翻译也带来了一些新的研究问题，例如模型过拟合问题 (Belinkov 等, 2018; Zhao 等, 2018; Vaibhav 等, 2019)，可解释性问题 (Ding 等, 2017; Strobelt 等, 2019; Voita 等, 2019) 和非自回归机器翻译 (Gu 等, 2018; Shao 等, 2019; Zhou 等, 2020) 等。

过拟合（Over-fitting）是神经网络模型普遍存在的一个基本问题，是指模型对训练集的拟合程度远高于对同分布测试集的拟合程度的现象。为了缓解过拟合问题，研究者们提出一些方法来降低模型的复杂程度或干扰模型对训练集的拟合，这类方法被称为正则化方法（Regularization Methods）。对于神经机器翻译而言，模型存在着词向量参数冗余度高、对噪声的鲁棒性差和对训练标签过度自信等具体问题，针对上述问题，研究者们分别提出了词向量正则化、扰动正则化和标签平滑等正则化方法。在后续的章节中，本文对上述问题进行了进一步的探究并提出了相应的正则化方法。

1.2　研究现状

本节将从神经机器翻译的基本原理，基本模型和主要研究方向三个方面对神经机器翻译的研究现状进行介绍。

1.2.1　基本原理

给定一个句对 $\mathrm{X} = \{x_1, x_2, ...x_{\mathrm{I}}\}, \mathrm{Y} = \{y_1, y_2, ..., y_{\mathrm{J}}\}$，神经机器翻译使用链式法则建模在给定源端句子的情况下的目标端句子的概率：

$$p(\mathrm{Y}|\mathrm{X}, \theta) = \prod_{j=1}^{\mathrm{J}+1} p(y_j|y_{<j}, \mathrm{X}, \theta), \tag{1.1}$$

其中，θ 代表模型的参数，I 和 J 分别为源端句子和目标端句子的长度，y_0 和 $y_{\mathrm{J}+1}$ 分别代表 BOS 和 EOS，即句子的起始符和结束符（Beginning and Ending of the Sentence）。

在这个基础上，神经机器翻译通过最大化训练数据的似然（Likelihood）来训练模型，具体来说是通过梯度下降（Gradient Descent）最小化训练集 D 中所有

样本的交叉熵（Cross Entropy Loss）：

$$\ell(\mathrm{X}, \mathrm{Y}, \boldsymbol{\theta}) = -\log p(\mathrm{Y}|\mathrm{X}, \boldsymbol{\theta}) \tag{1.2}$$

$$\mathcal{L} = \mathcal{L}(\mathrm{D}, \boldsymbol{\theta}) = \frac{1}{\mathrm{D}} \sum_{(\mathrm{X},\mathrm{Y})\in\mathrm{D}} \ell(\mathrm{X}, \mathrm{Y}, \boldsymbol{\theta}), \tag{1.3}$$

其中训练集 $\mathrm{D} = \{(\mathrm{X_n}, \mathrm{Y_n})\}_{\mathrm{n=1}}^{\mathrm{N}}$，N 为训练集 D 的大小。

1.2.2 基本模型结构

当前最主流的神经机器翻译架构是 Google 公司于 2017 年提出的 Transformer 模型 (Vaswani 等, 2017)，该模型沿用了 RNNSearch 模型 (Bahdanau 等, 2015) 的“编码器-解码器”架构（Encoder-Decoder Framework），并创新性地提出了自注意力机制（Self-Attention）和多头注意力机制（Multi-head Attention），不仅训练速度远超 RNNSearch 模型，而且在 WMT 英语-德语、英语-法语翻译任务上取得了历史最佳的翻译效果，是目前神经机器翻译研究的基线模型（Baseline）。

如图1.1 所示，Transformer 模型包含一个编码器（Encoder）和一个解码器（Decoder），编码器和解码器都由 N 层网络叠加组成。其中编码器的每一层由自注意力子层（Self-Attention ）和前馈神经网络子层（Feed Forward Network，FFN）组成，解码器的每一层由自注意力子层、交叉注意力子层（Cross-Attention）和前馈神经网络子层组成。Transformer 的每个子层都包含一个残差连接层（Residual Connection) 和一个规范化层 (Layer Normalization)。

Transformer 模型的自注意力和交叉注意力都采用了缩放点积注意力机制（Scaled Dot-Product Attention）。如图1.2（左）所示，缩放点积注意力由查询向量 Q、键向量 K、值向量 V 共同计算：

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}(\frac{\mathrm{QK^T}}{\sqrt{d_{\mathrm{model}}}})\mathrm{V}, \tag{1.4}$$

其中，d_{model} 是模型隐状态的维度。为了增强注意力机制的表达能力，Transformer 进一步提出了多头注意力机制（Multi-Head Attention）。如图1.2（右）所示，多头注意力机制先通过不同的线性映射分别将 Q、K 和 V 映射到 h 个子空间（每个子空间维度为 $\frac{d_{\mathrm{model}}}{h}$），然后在每个子空间独立地进行注意力计算，最后将结果拼接、映射，得到多头注意力的最终结果：

$$\begin{aligned} \mathrm{MultiHead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) &= \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)\mathrm{W^O} \\ \mathrm{head}_i &= \mathrm{Attention}(\mathrm{QW^Q}_i, \mathrm{KW^K}_i, \mathrm{VW^V}_i), \end{aligned} \tag{1.5}$$

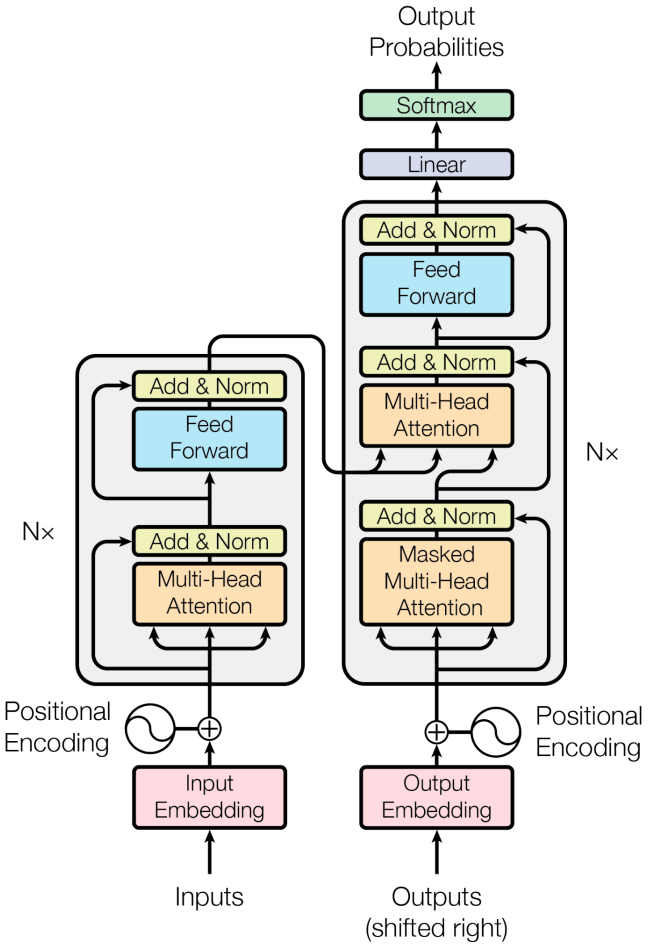


图 1.1 Transformer 的模型结构 (Vaswani 等, 2017)

Figure 1.1 The architecture of Transformer (Vaswani 等, 2017)

其中，$\mathrm{W}^{\mathrm{Q}}{}_i$、$\mathrm{W}^{\mathrm{K}}{}_i$、$\mathrm{W}^{\mathrm{V}}{}_i$ 分别是 Q、K、V 面向第 i 个子空间的映射矩阵，W^{O} 是多头注意力的输出映射矩阵。

Transformer 模型的训练高度并行，这使得它不能像循环神经网络（Recurrent Neural Network，RNN）那样通过循环读取输入词来建立时序信息，而是将输入词的位置信息作为相位，通过正弦和余弦函数建立固定的位置编码（Positional Encoding）：

$$\begin{aligned} \mathrm{PE}_{(\mathrm{pos},2i)} &= \sin(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}) \\ \mathrm{PE}_{(\mathrm{pos},2i+1)} &= \cos(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}), \end{aligned} \tag{1.6}$$

其中，pos 为输入词的在句子中的序号，$2i$ 和 $2i+1$ 为位置编码的维度。在输入层，位置编码被添加到输入词的词向量上，参与后续的模型计算。

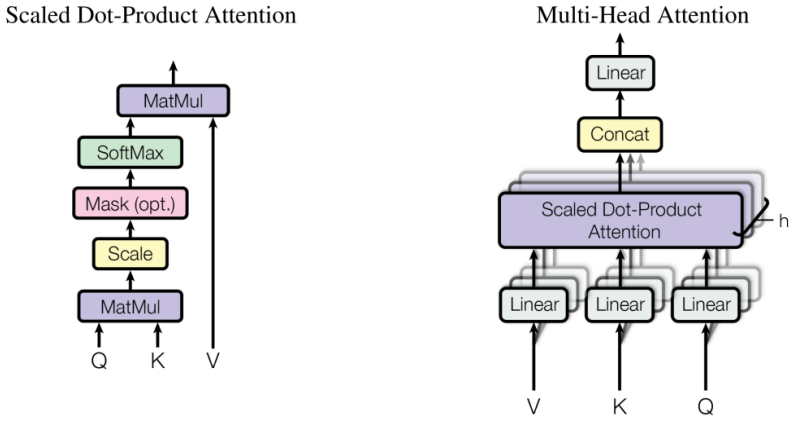


图 1.2 Transformer 的注意力机制 (Vaswani 等, 2017)

Figure 1.2 Attention Mechanism of Transformer (Vaswani 等, 2017)

1.2.3 机器翻译的性能评价指标

机器翻译的性能一般是指机器翻译模型生成译文的好坏。衡量译文准确与否的最佳方式是通过多名语言专家进行人工评价，该评价涉及译文的忠实度、流畅度、充分度等多个维度。然而人工评价效率低、成本高，具有极大的主观性，无法满足学术研究场景下的评价需求，因此研究人员提出了一些自动化评价指标，基于人工翻译译文来评价机器翻译译文的好坏，其中最常用的是基于 n 元文法匹配的混合精确度指标 BLEU：

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{\text{N}} w_n \log p_n) \tag{1.7}$$

其中 w_n 和 p_n 表示 n 阶文法匹配的权重和精确度，BP 为长度惩罚因子（Brevity Penalty），计算方式为：

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{1.8}$$

其中 c 和 r 分别为系统译文长度和参考译文的长度。

1.2.4 主要研究方向

当前，神经机器翻译的前沿探究大致分为以下几个方向：

- 低资源机器翻译。神经机器翻译的优异性能一方面依赖神经网络强大的拟合能力，另一方面也依赖充足的平行语料。在训练语料较少的情况下，神经机器翻译模型的性能会出现严重的下降。然而，世界上绝大多数语言使用人口较

少，相关的平行语料较为稀缺，因此研究者们提出了一些方法以改进低资源条件下神经机器翻译的性能。Sennrich 等 (2016a) 提出反向翻译（Back Translation）的方法扩充训练语料，该方法借助已有的平行语料训练目标语言到源语言的反向翻译模型，然后利用该模型将目标语言的单语语料翻译成源语言，生成大规模的伪平行语料，用于训练源语言到目标语言的机器翻译模型。Lample 等 (2018) 提出一种基于自编码器的无监督机器翻译方法，其中编码器将句子映射到中间语义空间的隐变量，解码器将隐变量重构回原始的句子。该方法训练两种语言的编码器将源语言句子和目标语言句子映射到同一个语义空间，然后利用源语言的编码器和目标语言的解码器实现了源语言到目标语言的翻译。由于语言交流的局域性，有时两种语言的平行语料较少，而它们与第三种语言的平行语料较多。为此，Cheng 等 (2017) 提出以中间语言为媒介，分别训练源语言到中间语言和中间语言到目标语言的翻译模型，将两个模型串联以实现源语言到目标语言的机器翻译。

- 非自回归机器翻译。普通的神经机器翻译学习在给定源端句子和目标端前缀的情况下生成下一个词，这导致模型在解码时需要采取自回归（Autoregressive）的方式一个词一个词地串行解码，模型的解码速度受到很大的限制。为此，Gu 等 (2018) 提出非自回归（Non-autoregressive）神经机器翻译，在给定源端句子的条件下，对目标端句子的每个词进行独立的建模，使得模型可以并行解码出整个目标端句子。非自回归机器翻译虽然解码速度很快，但翻译效果较差，存在比较严重的过翻译和漏翻译的问题。为此，研究者们对非自回归机器翻译进行了不断的研究和改进。研究者们发现，序列级别的知识蒸馏 (Kim 等, 2016) 能极大地提升非自回归机器翻译的性能。另外，Wei 等 (2019) 提出通过模仿学习（Imitation Learning）的方式让非自回归模型学习自回归模型的动作序列；Guo 等 (2020b) 提出让非自回归模型与自回归模型共享参数，通过课程学习的方式使得模型的训练逐渐从自回归过渡到非自回归；Li 等 (2019b) 提出利用自回归模型中的隐状态和注意力的提示（Hint）引导非自回归模型的训练。

- 多语言机器翻译。普通的机器翻译模型只能实现一对一的语言翻译，这样有两个问题：（1）如果要构建一个多对多的翻译系统，每个翻译方向都需要训练一个单独的翻译模型，这样系统的训练成本和使用成本比较高；（2）一些低资源的翻译方向需要通过中间语言进行转接翻译，存在错误积累的问题。为此，研

究者们提出通过单个模型实现多对多的机器翻译来减少成本并提升低资源方向的翻译质量。Dong 等 (2015) 提出了一个从单种源语言到多种目标语言的翻译模型，该模型包含一个编码器和多个相互独立的解码器，其中每个解码器对应一种目标语言。Luong 等 (2016) 提出了一个从多种源语言到多种目标语言的翻译模型，该模型包含多个编码器和多个解码器，其中每个编码器对应一种源语言，每个解码器对应一种目标语言。Ha 等 (2016) 对源语言句子的每个词添加指示源语言的符号，对源语言句子的两端添加目标语言的符号，使用一个编码器和一个解码器实现了多对多的机器翻译系统。

- 同声传译。在国际会议的场景中，等待发言人说完整句话再进行翻译会造成翻译的较大延迟，降低参会人员的交流体验和效率。同声传译模型能够在仅读入部分源语言句子的情况下开始输出译文，并随着源语言句子的继续读入不断补充和完善译文。Cho 等 (2016) 首次提出了基于神经机器翻译的同声传译模型，该方法根据模型翻译概率的变化动态地制定读写策略。Gu 等 (2017) 以翻译质量指标 BLEU 和翻译延迟指标为奖赏值，使用强化学习算法让模型自动地学习读写策略。Ma 等 (2019); Dalvi 等 (2018) 提出了固定延迟的同声传译模型，该方法有输出速度稳定的优点，但无法根据实际语境动态地调整读写速度。Zheng 等 (2019a,b) 提出对同声传译的读写策略进行监督学习，使得模型能够更加合理地执行读写决策。

- 多模态机器翻译。随着现代媒体的发展，机器翻译越来越多地出现在多模态的场景中，例如在对视频字幕进行翻译的时候，视频内容通常与字幕相对应，可以为字幕的翻译提供参考。目前多模态机器翻译一般以图像为额外的模态信息辅助文本的翻译。Calixto 等 (2017) 将卷积神经网络编码的视频信息添加到机器翻译模型的输入中或者用其初始化模型的隐状态来辅助文本翻译。Caglayan 等 (2016) 提出多模态的注意力机制，模型能够同时对图像和文本做注意力，充分地获取和利用图像信息用于辅助文本翻译。Ive 等 (2019) 将模型的解码分解为两个步骤，模型在第一步基于文本信息生成译文草稿，在第二步加入图像信息对译文进行完善。

- 篇章机器翻译。在翻译整篇文章时，机器翻译模型会独立地翻译每个句子。然而篇章的内容通常上下相关，忽略句子之间的联系可能会导致翻译错误。篇章机器翻译能够利用篇章中的上下文信息来辅助当前句子的翻译。Jean 等 (2017) 基

于循环神经网络机器翻译模型实现了对上下文信息的有效利用，该方法使用额外的网络模块对上下文信息进行编码并用于当前句的机器翻译。Voita 等 (2018) 基于 Transformer 实现了篇章翻译，该方法通过两个编码器分别对上下文和源语言句子进行编码，两部分信息通过门控机制融合后被输出到模型的解码器进行解码。Zhang 等 (2018) 在翻译模型的编码器和解码器加入了对上下文的注意力模块，使得模型能够通过注意力机制收集上下文信息来辅助当前句的翻译。

- 领域自适应。领域外的机器翻译模型通常在领域内表现得很差，而领域内的语料通常是稀缺的，因此如何使领域外的机器翻译模型适应领域内的翻译是一个重要的研究问题。一些方法通过筛选领域外的语料来扩充训练数据。Moore 等 (2010) 使用领域内和领域外的语言模型对领域外的语料进行打分并将分差较小的语料加入训练集；Wang 等 (2017) 使用神经机器翻译模型的句级别向量对领域外的语料与领域内的语料进行对比，选用其中与领域内相似的语料。另外，一些方法通过改进训练方法提升模型的领域自适应能力。Freitag 等 (2016) 先在所有训练语料上训练机器翻译模型，然后使用领域内的语料对模型进行微调;Chu 等 (2017) 在微调时加入部分领域外语料，通过混合微调的方式减轻模型对领域外语料的知识遗忘；Barone 等 (2017) 在微调时添加正则项以减少模型对领域内语料的过拟合。

- 模型正则化。神经机器翻译模型存在容易过拟合的缺点，为此研究者们提出很多正则化方法以减轻模型的过拟合。最通用的正则化方法是 Dropout (Hinton 等, 2012) 和标签平滑 (Szegedy 等, 2016)，它们被广泛应用于包括神经机器翻译在内的各种任务。对于神经机器翻译，研究者针对一些具体问题提出了一系列正则化方法。例如针对子词分割问题，Kudo (2018a) 提出从多个可能的子词候选中采样，Park 等 (2020a) 提出使用对抗性的子词分割；针对噪声鲁棒性问题，Wu 等 (2019) 和 (Miyato 等, 2017) 分别提出在输入中添加词删除噪声和对抗性噪声以提升模型对噪声的鲁棒性；针对模型的注意力机制，Zhang 等 (2019a)、You 等 (2020) 和Li 等 (2018) 分别提出相应的正则化方法来促进注意力分布的稀疏性、局部性和多样性。

1.3 本文研究问题

过拟合问题（Over-fitting）是神经网络模型普遍存在的一个基本问题，是指模型对训练集的拟合程度远高于对同分布测试集的拟合程度的现象。过拟合的主要原因是神经网络模型的拟合能力很强，而模型的训练损失函数只考虑了对训练集的拟合，因此模型往往过于紧密地拟合已知的训练集而忽略了在其它数据上的表现。为了缓解神经网络模型的过拟合，研究者们提出一些方法来降低模型的复杂程度或干扰模型对训练集的拟合，这类方法被称为正则化方法（Regularization Methods）。

研究神经网络模型的正则化方法具有重要意义。一方面，正则化方法能让模型的参数得到更高效的利用，降低模型的训练成本和使用成本；另一方面，通过对正则化方法的研究，我们可以进一步探究神经网络模型的工作原理，为模型的结构设计和训练方式的改进提供借鉴和启发。

本文对词向量正则化、扰动正则化和标签平滑这三类正则化问题进行了进一步的探究。这三类正则化方法分别针对模型复杂度过高、训练数据相对稀缺和训练标签过于武断这三个引起过拟合的具体原因，提出通过限制模型参数的学习、对训练数据加噪声和标签平滑的方式来缓解过拟合问题。本文接下来对这三个研究问题进行简要的介绍。

1.3.1 词向量正则化

表 1.1 权重衰减在 WMT 英文-罗马尼亚文数据集上的实验结果

Table 1.1 Experimental results of Weight Decay on WMT En-Ro task

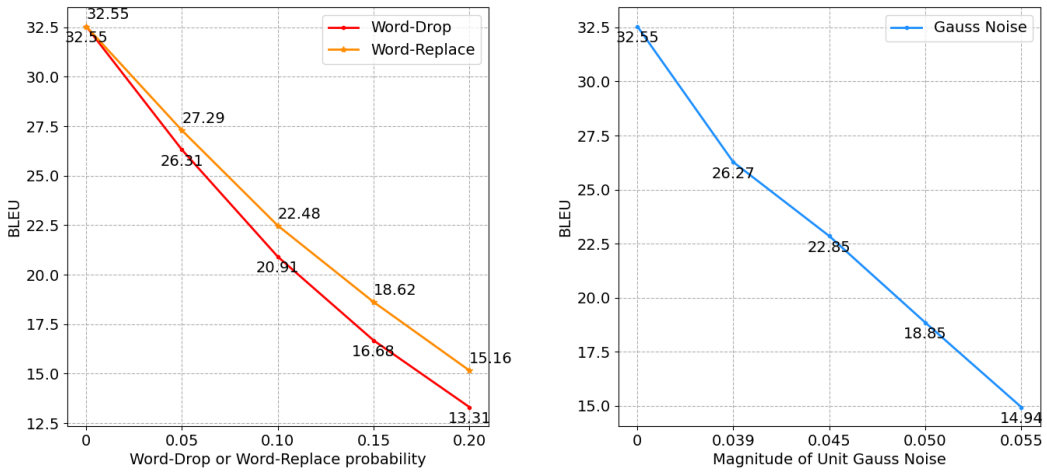
Configuration	BLEU
w/o Weight Decay	32.58
+ Weight Decay = 2e-6	32.53
+ Weight Decay = 1e-5	32.59
+ Weight Decay = 1e-4	32.26
+ Weight Decay = 1e-3	31.78

模型复杂度高是神经机器翻译模型容易过拟合的原因之一，一些正则化方法通过约束模型的参数来降低模型的复杂度并缓解过拟合。词向量正则化是作用于模型的词向量（Word Embedding）参数的正则化方法，对于解决神经机器翻译模型的过拟合问题十分重要。首先，神经机器翻译模型输入层的词向量参数占

总参数的 37.5% 左右，是模型过拟合的重要来源；其次，相比其它参数，词向量参数的可解释性更好，词向量正则化的可行性更高。

约束参数学习的最基础的方法是权重衰减，然而如表 1.1 所示，权重衰减对神经机器翻译模型没有明显的提升。近年来，一些工作通过共享参数或添加促进词对齐的正则项来约束词向量参数的学习并取得了一定的效果 (Press 等, 2017; Kuang 等, 2018; Liu 等, 2019)。然而，这些方法通常依赖相同字符系统下的词形匹配或者外部对齐工具的辅助来实现，在应用范围和使用的灵活性方面有较大的局限性。总体而言，目前对词向量参数进行正则化的研究仍然较少，已有的工作存在着一些较为明显的缺陷，词向量正则化仍然有较大的改进空间。

1.3.2　扰动正则化



(a)　　**(b)**

图 1.3 （a）模型对词删除、词替换的鲁棒性；（b）模型对高斯噪声的鲁棒性。

Figure 1.3 (a) The robustness of model to Word-Dropout and Word-Replacement noises. (b) The robustness of model to Gaussian noises.

神经机器翻译模型过拟合的表现之一是对噪声的鲁棒性很差。如图1.3所示，在输入中添加 10%-20% 的词删除或词替换噪声或者 4.5%-5.5% 的高斯噪声就会使得神经机器翻译模型的性能出现大幅度的下降。为此，研究者们提出扰动正则化（Perturbation Regularization）方法，该类方法通过在训练样本中添加少量的噪声以提升模型对噪声的鲁棒性。已有的扰动正则化方法专注于寻找更好的扰动形式，提出了词删除 (Gal 等, 2016)、词替换 (Bengio 等, 2015; Wu 等, 2019)、对

抗性扰动等 (Miyato 等, 2017; Sato 等, 2019b) 不同的扰动方法。然而实验结果证明，已有的扰动正则化方法对模型性能的提升较为有限，而且不同类型的扰动带来的提升差异不大。目前扰动正则化方法在方法原理和训练机制方面还需要更深层次的探索。

1.3.3 标签平滑

神经机器翻译使用硬标签（Hard Label）训练模型，这使得模型容易过拟合训练集中出现的标签而忽略掉其它可能正确的标签。为了解决这个问题，Szegedy 等 (2016) 提出标签平滑（Label Smoothing）方法，使用平均分布标签来平滑硬标签，以缓解神经网络模型对硬标签过度自信（Over Confident）的问题。该方法简单有效且适用于各种分类任务，但其先验假设显然不是最优的。另一种有效的标签平滑方法是知识蒸馏 (Hinton 等, 2015)。其中自蒸馏 (Furlanello 等, 2018) 使用训练好的同结构模型来蒸馏重新初始化的模型，是一种简单有效的标签平滑方法。但是自蒸馏具有训练代价大、标签生成器过于臃肿和不能在线使用的缺点。神经机器翻译模型需要一种高效的在线标签平滑方法。

1.4 主要研究内容

本文对词向量正则化、扰动正则化和标签平滑三类正则化方法进行了进一步的探索研究，并针对它们各自存在的问题分别提出了相应的解决方案。

1.4.1 基于语义相关性的词向量正则化方法

为了更好地规范神经机器翻译模型的词向量参数，本文提出一种能够增强词向量语义相关性的词向量正则化方法，该方法完全基于双语语料的内在特征，适用于所有翻译方向且不依赖外部的知识指导。具体而言，机器翻译的一个训练样本由一个源端句子和一个目标端句子组成，其中隐含了单语言的词共现信息和双语言之间的词共现信息。基于这两种共现信息，本文提出一个自编码的训练目标来同时促进词向量的单语相关性和对齐相关性。本文的方法不增加参数，能够与神经机器翻译模型共同训练。实验结果表明，本文的方法能够极大地增强词向量参数的语义特征，并显著提升神经机器翻译模型的性能。

1.4.2　基于预测差异的扰动正则化方法

扰动正则化方法通过对训练数据加噪声以提升模型对噪声的鲁棒性，其隐含的前提是模型对原始样本的拟合程度总是高于对被扰动样本的拟合程度，但该前提并不总是成立。本文提出利用模型在输入受到扰动前后对目标端词的预测差异来分析词级别样本的拟合情况，发现模型对相当一部分原始样本是相对欠拟合的，对这部分样本进行扰动正则化训练会损害模型的性能，这说明已有扰动正则化方法的训练模式不是最优的。为了同时缓解模型的过拟合和欠拟合问题，本文提出将输入扰动引起的预测差异作为额外的正则项来训练神经机器翻译模型。在常用的数据集上，本文的方法相比现有的方法取得了巨大的提升，并显著地改善了模型对噪声的鲁棒性。

1.4.3　基于条件变分自编码器的标签平滑方法

本文提出一种轻量、高效的在线标签平滑方法以缓解模型对训练集标签过度自信的问题。本文利用条件变分自编码器能够基于给定条件将数据标签编码到一个隐变量空间并重构出来的特点，为神经机器翻译设计了一个基于条件变分自编码器的平滑标签生成器。该生成器能够学习给定源端和目标端输入条件下的词级别标签的隐变量分布，并通过该分布实时采样隐变量并生成新标签，用于神经机器翻译模型的在线标签平滑。实验结果表明，本文的方法能够较好地缓解模型对训练标签过度自信的问题并提升模型的翻译性能。

1.5　章节组织

本文的章节组织结构如下：

第一章介绍了神经机器翻译的研究背景和意义，发展现状、基本原理和主流模型，简要概括了当前神经机器翻译的前沿方向，以及本文研究的主要问题与主要内容。

第二章介绍了词向量正则化方法的相关研究，提出基于语义相关性的词向量正则化方法，并分析了词向量的语义特征和空间分布特征。

第三章介绍了扰动正则化方法的相关研究，提出利用模型在输入被扰动前后对目标词的预测差异来分析词级别样本的拟合情况，并提出将输入层扰动引起的预测差异作为额外的正则项以同时缓解模型的过拟合和欠拟合问题。

第四章介绍了利用条件变分自编码器进行标签平滑的研究，介绍了条件变分自编码器的基本原理，提出利用条件变分自编码器构建平滑标签生成器。

第五章对所有工作进行了总结与展望。

第 2 章　基于语义相关性的词向量正则化方法

2.1　引言

神经网络模型通常具有巨大的参数量，这使得模型具备很强的拟合能力，但也导致了模型容易过拟合的问题。为了减轻神经网络模型的过拟合，研究者们通常使用 Dropout (Hinton 等, 2012) 和权重衰减 (Krogh 等, 1991) 等正则化方法来限制参数的学习。

对于机器翻译任务，神经网络模型还包含一类特殊的参数——词向量参数。在模型中，源端和目标端语言词表中的每一个词都对应着一个参数向量，这些参数向量和其它参数一样通过梯度下降的方式更新学习。近年来，一些作用于神经机器翻译模型词向量参数的正则化方法被提出并取得了良好的效果，本章在以上工作的基础上，针对词向量参数正则化进行了进一步的探究。

词向量参数的正则化对于神经机器翻译模型具有特殊的重要性，主要体现在以下两个方面：

首先，神经机器翻译模型涉及两种不同的语言，每个语言的词表通常包含 3-4 万个词，词表的参数向量占模型参数的主要部分，因此词向量参数的正则化对于整个模型的表现来说至关重要。例如对于 NIST 中文-英文机器翻译模型来说，源端输入层词向量（Encoder Input Embeddings）、目标端输入层词向量（Decoder Input Embeddings）和目标端输出层词向量（Decoder Output Embeddings）（类别特征向量）三者占模型总参数量的 53.4%，其中源端和目标端输入层词向量参数占总参数量的 37.5%。

其次，与神经机器翻译模型的其它参数不同，词向量参数将词表的每个词对应到欧几里得空间的一个向量，具有较好的可解释性，使得研究者们更容易设计出有针对性的正则化方法，因此词向量正则化具有更高的可行性。如表 1.1，对神经机器翻译模型的所有参数采取 L_2 正则化没有取得明显的效果，然而一些面向词向量参数的正则化工作却取得了显著的效果 (Press 等, 2017; Yang 等, 2019; Kuang 等, 2018)。

本章希望通过规范词向量参数的语义相关性，即通过自编码的方式使具有相近语义的词具有相近的词向量参数来提升神经机器翻译模型的性能。该想法

主要出于以下三个动机：

1. 神经机器翻译模型的词向量参数具有较高的冗余度。例如 Press 等 (2017); Liu 等 (2019) 等通过参数共享的方式将模型的词向量参数减少了一半左右，然而模型的翻译性能没有受到明显的影响。

2. 词向量参数缺乏对齐相关性的约束。在神经机器翻译模型中，源端和目标端的词向量参数位于模型的两个终端，两者之间没有显式的连接，梯度下降的参数更新方式无法使两者学习到明显的对齐关系。

3. 词向量参数缺乏单语相关性的约束。神经机器翻译模型的训练目标同样缺乏对单语相关性的显式约束，因此难以保证词向量参数能够学习到单语间的语义相关性，对于缺少单向掩码和语言模型任务的编码器来说尤其如此。

近年来，一些促进词向量参数之间的相关性的方法被陆续提出，这些方法取得了一定的效果，但也分别存在一些问题。Press 等 (2017) 提出通过共享源端和目标端词表的方式减少词向量参数的冗余度并提高对齐相关性，具体做法是让源端词表和目标端词表中具有相同词形（Word form）的词使用同一个参数向量。然而该方法仅适用于那些使用相同字符系统的语言对，不能应用于英文和日文这种字符系统完全不同的语言对，而后者是更为普遍的情况。另外，两种语言中具有相同词形的词可能具有完全不同的含义。例如日语中“丈夫”的含义是坚固、结实，而汉语中“丈夫”通常是指已婚女子的配偶，两者使用同一个参数向量可能会导致模型产生错误的翻译。Kuang 等 (2018) 提出借助源端和目标端之间的注意力权重来确定对齐关系，并拉近对齐词向量之间的距离。然而 Transformer 模型经过多层的自注意力机制之后，每个词的高层特征都融合了整个句子的信息，高层注意力权重得到的对齐关系并不准确。Liu 等 (2019) 提出通过外部词对齐工具得到词对齐关系并共享对齐词的部分参数，以增强两者的对齐相关性。该方法取得了较为显著的效果，但也存在以下缺点：（1）词对齐关系依赖外部的词对齐工具获得；（2）需要人工设置词向量参数共享的比重，难以实现对每个词的准确调整；（3）只提升了对齐相关性，无法提升单语相关性。（4）通过大比例参数共享的方式增强对齐性，让占 30% 的没有明显对齐关系的词之间也共享参数，可能会降低模型的表达能力。

本章提出利用双语语料的内在特征来促进词向量参数的语义相关性。具体而言，机器翻译的一个训练样本由一个源端句子和一个目标端句子组成，其中

隐含了单语言的词共现信息和双语言之间的词共现信息。基于这两种共现信息，本章提出一个自编码的训练目标来同时促进词向量的单语相关性和对齐相关性。具体实现上，本章借鉴 Skip-gram 方法的训练目标并对其进行了以下拓展：（1）将单语的应用场景拓展到双语；（2）将训练方式拓展到句级别的并行训练，使其能够和神经机器翻译模型共同训练；（3）方法中的词向量参数和类别映射矩阵都使用机器翻译模型的输入层词向量参数代替，无需增加额外的参数。本章的方法不依赖外部的知识指导，不需要增加参数，实现方便，具有较低的训练成本。

本章在 NIST 中文-英文数据集上对方法进行了实验。实验结果表明，本章的方法能够显著提升神经机器翻译模型的性能，相比基线模型提升了 1.69 个 BLEU 值，相比使用外部对齐工具的 Shared-private 方法提升了 0.59 个 BLEU。对词向量参数的分析实验表明，本章的方法显著地增强了词向量参数的语义相关性，有效地促进了源端和目标端词向量空间的融合。

2.2　相关工作

本章的方法对神经机器翻译模型的词向量参数进行了增强语义相关性的自编码训练，这主要涉及到三个方面的内容：普通的词向量正则化方法、词向量的单语相关性、词向量的对齐相关性。本节将分别对这三个方面的相关工作进行简要的介绍。

2.2.1　普通的词向量正则化

近年来在自然语言处理领域陆续出现了一些作用于模型词向量的正则化方法，这里简要介绍其中不涉及语义相关性的工作。Berend (2017, 2018) 分别在词性标注（Part-of-speech Tagging）、命名实体识别（Named Entity Recognition）和多词表达识别（Multi-Word Expression Identification）任务上探究了对模型的词向量参数进行 L_1 正则化的效果，发现 L_1 词向量正则化对于这些任务有着显著的提升。Demeter 等 (2020) 发现使用交叉熵训练的神经语言模型存在词向量分布不均衡的问题，低频词的词向量长度更小，一些低频词分布在词向量空间凸包的内部，无法被解码出来。Meng 等 (2019) 提出了球面词向量的解决方法，该方法将所有词向量的长度全部限制为 1 以解决数据不平衡造成的词向量分布不平衡的问题。Unanue 等 (2019) 为神经机器翻译模型增加了一个预测目标端输入层词向量的目标函数以增强模型的泛化性能。

2.2.2 词向量的单语相关性

在自然语言处理领域，一些方法通过自编码的方式从单语语料中学习具有语义信息的词向量，包括轻量的 Word2vec (Le 等, 2014; Mikolov 等, 2013)、Skip-thought (Kiros 等, 2015)、Glove(Pennington 等, 2014) 等以及近年来备受关注的大型预训练语言模型 BERT (Devlin 等, 2019)、GPT (Radford 等, 2018) 等。另外，一些工作通过将词划分为更细粒度的子词或字符来间接地增强词之间的单语相关性。Sennrich 等 (2016b); Johnson 等 (2017); Ataman 等 (2018) 提出将词划分为子词（subword），Cho 等 (2014); Ling 等 (2015); Costa-jussà 等 (2016); Chen 等 (2016) 采用了字符级别的机器翻译系统。

2.2.3 词向量的对齐相关性

近年来，一些工作提出增强神经机器翻译模型词向量的对齐性以提高神经机器翻译的性能。Press 等 (2017) 提出合并源端和目标端的词表并让两端具有相同词形的词共享同一个参数向量，这样具有相同词形的两个词就建立了对齐关系。Yang 等 (2019) 提出将模型编码器输出的句级别向量和解码器输入层的句级别向量的 L_2 距离作为正则项以实现句级别的对齐。Kuang 等 (2018) 以编码器和解码器的注意力权重作为对齐依据，拉近对齐源端词和目标端词的距离。Liu 等 (2019) 利用词对齐工具获取词对齐关系，共享对齐词的词向量的部分维度，将剩余未共享的维度作为独立学习的参数，这样同时保证了对齐词之间的共性和特性。Mi 等 (2016); Liu 等 (2016); Cheng 等 (2016); Feng 等 (2017) 使用外部的词对齐知识来辅助机器翻译，Garg 等 (2019) 利用统计工具获取的词对齐知识监督训练神经机器翻译模型学习词对齐信息，Li 等 (2019a) 利用预测差异从复杂的神经网络机器翻译模型中提取词对齐信息。

2.3 研究背景

本节介绍 Word2vec (Mikolov 等, 2013; Le 等, 2014) 的基本原理，以及衡量对齐相关性的指标——对齐错误率。

2.3.1 Word2vec 的基本原理

在深度学习被大规模应用于自然语言处理之前，自然语言处理任务的模型通常使用词表中的序号代表每个词，将每个词看作一个独立的符号，这种做法无

法体现出词与词之间的相关性。为了能更好地表征词语之间的相关性，Mikolov 等 (2013); Le 等 (2014) 提出 Word2vec 方法，使用欧几里得空间的向量表示每个词，并利用大量语料中的词共现信息来训练词向量。

Word2vec 的基本原理是两个词的相关性与它们共现于同一个句子中的频率正相关。例如“总统”和“访问”两个词具有较强的相关性，经常出现在同一个句子中；而“键盘”和“雾霾”两个词基本没有相关性，很少出现在同一个句子中。

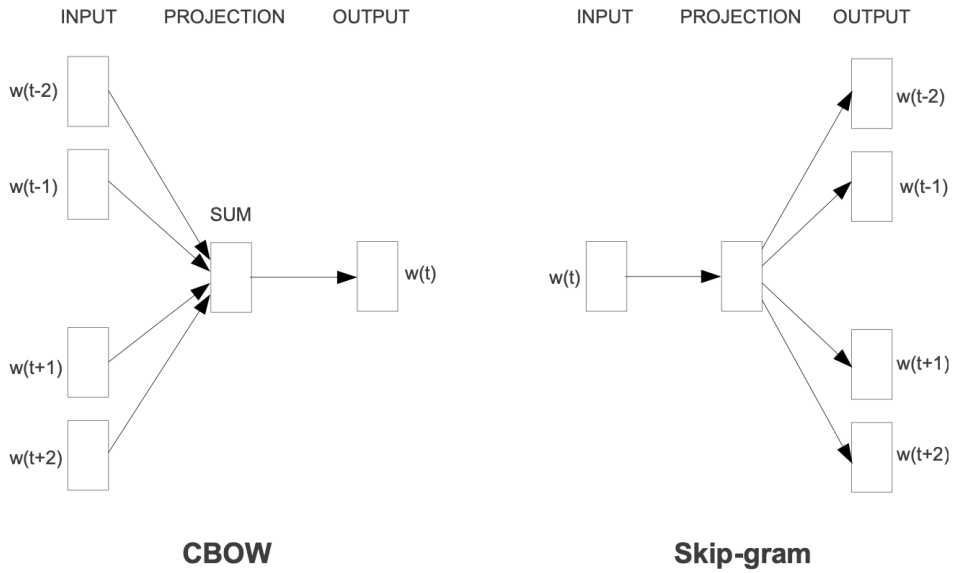


图 2.1 CBOW 和 Skip-gram 的结构示意图 (Mikolov 等, 2013)

Figure 2.1 Architectures of CBOW and Skip-gram(Mikolov 等, 2013)

对于一个长度为 I 的句子 W = [w_1, w_2, ..., w_{I-1}, w_I]，Word2vec 让其中的每个词轮流作为中心词，通过自编码的形式增强中心词与其窗口中其它词（上下文）的相关性。Word2vec 包含两种具体的方法，分别是连续词袋模型（Continuous Bag-of-Words Model, CBOW）和连续跳词模型（Continuous Skip-gram Model, Skip-gram），图2.1展示了两个模型的基本框架。

其中，CBOW 的训练目标是最大化上下文预测中心词的概率：

$$\mathbf{c}_t = \text{average}(\mathbf{e}_{t-k}, ..., \mathbf{e}_{t-1}, \mathbf{e}_{t+1}, ..., \mathbf{e}_{t+k}), \tag{2.1}$$

$$p(w_t|w_{t-k}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k}) = \frac{\exp(\tilde{\mathbf{e}}_t^{\mathrm{T}}\mathbf{c}_t)}{\sum_m^{|\mathrm{V}|}\exp(\tilde{\mathbf{e}}_m^{\mathrm{T}}\mathbf{c}_t)}, \tag{2.2}$$

$$l = -\frac{1}{\mathrm{I}}\sum_{t=1}^{t=\mathrm{I}}\log p(w_t|w_{t-k}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k}), \tag{2.3}$$

其中 k 是窗口的大小，$\mathbf{e}_{t-k}$ 是窗口词 w_{t-k} 的输入层词向量，$\mathbf{c}_t$ 代表上下文信息，是所有窗口词的词向量的平均，$\tilde{\mathbf{e}}_t$ 是中心词 w_t 的输出层词向量，V 是模型的词表。

与 CBOW 相反，Skip-gram 的训练目标是最大化中心词预测上下文的概率：

$$p(w_c|w_t) = \sum_{g=t-k,...,t-1,t+1,...,t+k}\frac{\exp(\tilde{\mathbf{e}}_g^{\mathrm{T}}\mathbf{e}_t)}{\sum_m^{|\mathrm{V}|}\exp(\tilde{\mathbf{e}}_m^{\mathrm{T}}\mathbf{e}_t)}, \tag{2.4}$$

$$l = -\frac{1}{\mathrm{I}}\sum_{t=1}^{t=\mathrm{I}}\log p(w_c|w_t), \tag{2.5}$$

其中 k 是窗口的大小，w_c 指上下文，w_g 是上下文中的某个具体的词，$\tilde{\mathbf{e}}_g$ 是 w_g 的输出层词向量，$\mathbf{e}_t$ 是中心词 w_t 的输入层词向量。

2.3.2 对齐错误率

对齐错误率（Alignment Error Rate, AER）(Och 等, 2000) 是衡量词对齐质量的指标，一般通过对比模型的词对齐和人工标注的词对齐计算得到。然而，人工标注词对齐是一项复杂的工作，因为两种不同语言的语法和词汇通常无法完全对应。对于英文和中文的词对齐来说，英文中的冠词、介词和助动词等通常起到辅助作用，没有具体的语义，很难对齐到合适的中文词。例如在平行语对（“东北是国家的粮仓”，“The northeast is the granary of the country.”）中，英文句子中的三个冠词 “the” 都很难找到相应的中文对齐词。对于不明确的词对齐，Och 等 (2000) 提出使用让语言专家在标注时区分出两种词对齐：明确的词对齐（Sure, S）和不明确的词对齐（Possible, S），其中集合 S 包含于集合 P（$\mathrm{S} \subseteq \mathrm{P}$）。

在获得人工标注对齐数据的基础上，对于模型给出的词对齐集合 A，可以分别计算出其召回率（Recall）、精确率（Precision）和对齐错误率（AER）。其中，召回率衡量了模型词对齐的充分性：

$$\text{recall} = \frac{|\mathrm{A} \cap \mathrm{S}|}{|\mathrm{S}|}, \tag{2.6}$$

精确率衡量了模型词对齐的准确性：

$$\text{precision} = \frac{|\text{A} \cap \text{P}|}{|\text{A}|}, \tag{2.7}$$

对齐错误率同时考虑了召回率和精确率两个指标，用来衡量模型词对齐的整体质量：

$$\text{AER}(\text{S}, \text{P}; \text{A}) = 1 - \frac{|\text{A} \cap \text{S}| + |\text{A} \cap \text{P}|}{|\text{A}| + |\text{S}|}. \tag{2.8}$$

2.4　基于语义相关性的词向量正则化方法

神经机器翻译使用交叉熵损失和梯度下降算法训练模型，由此得到的词向量参数通常具有较差的语义特征。为了增强词向量参数的语义相关性，本章提出一种基于平行语料词共现信息的词向量自编码方法，简称为语义相关正则化。本章的目标是同时增强词向量的单语相关性和对齐相关性，其中单语相关性表示语义相关的两个源端词或两个目标短词应具有相近的词向量，对齐相关性表示两个对齐的源端词和目标端词应具有相近的词向量。

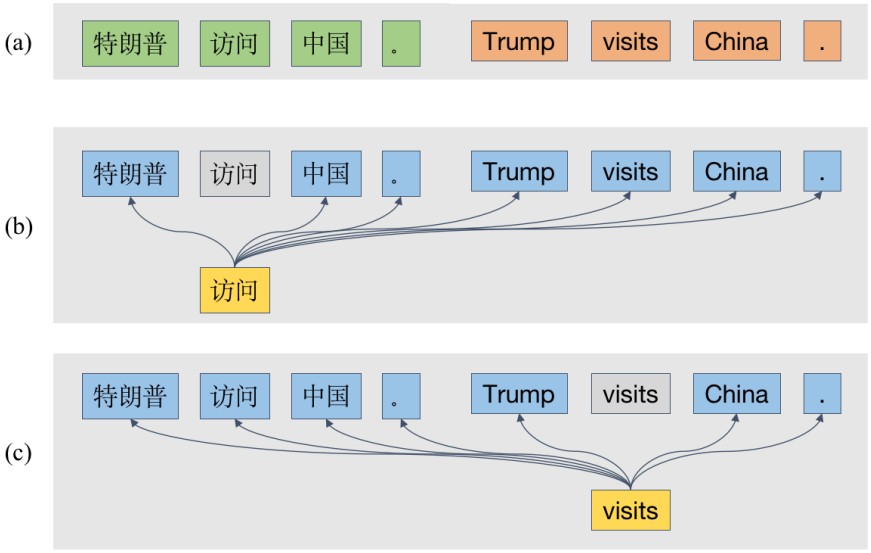


图 2.2 语义相关词向量正则化的示意图

Figure 2.2 Illustration for semantic regularization of word embedding

具体而言，如图 2.2 (a)，每个机器翻译的训练样本由一个源端句子和一个目标端句子组成，即源端句子 [“特朗普”，“访问”，“中国”，“。”] 和目标端句

子 [“Trump”, “visits”, “China”, “.”]。语义相关正则化将整个平行句对作为窗口，要求平行句对的词能够相互预测，即源端的每个词能够预测平行句对中除自身外的所有其他词，目标端的每个词也能够预测平行句对中除自身外的所有其他词。如图 2.2 (b)(c) 所示，当源端词 “访问” 为中心词时，需要预测其窗口词 [“特朗普” , “中国” , “。” , “Trump” , “visits” , “China” , “.”]; 当目标端词 “visits” 为中心词时，需要预测其窗口词 [“特朗普”, “访问”, “中国”, “。”, “Trump”, “China”, “.”]。

通过该训练目标，整个平行句对里的词能够建立起语义相关性，包括单语相关性和对齐相关性：（1）对于中文词 “访问” 和 “中国”，由于语料中频繁出现国外领导人访问中国或者中国领导人访问外国的新闻，因此这两个词也能够建立起单语相关性。（2）对于对齐词 “访问” 和 “visits”，由于它们频繁地出现在同一个平行句对，因此能够建立起对齐相关性。

语义相关正则化使用交叉熵损失训练词向量。对于平行句对 (X, Y)，其中每一个词的词向量构成词向量集合 $\mathrm{U} = \{u_i\}^{\mathrm{I+J}}$，I 和 J 分别是源端句子和目标端句子的长度。语义相关正则化的损失函数为词向量集合 U 中的每个词预测其它所有词的交叉熵：

$$p(u_j|u_i) = \frac{\exp(u_j^{\mathrm{T}} u_i)}{\sum_m^{|\mathrm{V}_s|+|\mathrm{V}_t|} \exp(e_m^{\mathrm{T}} u_i)}, \tag{2.9}$$

$$\ell_{\mathrm{emb}}(\mathrm{X}, \mathrm{Y}) = -\frac{1}{(\mathrm{I}+\mathrm{J})} \sum_i^{\mathrm{I+J}} \sum_{j \neq i}^{\mathrm{I+J}} \frac{1}{(\mathrm{I}+\mathrm{J}-1)} \log p(u_j|u_i), \tag{2.10}$$

其中 e_m 是词表中第 m 个词的词向量，V_s 和 V_t 分别是源端和目标端的词表。

训练集所有样本的正则化损失为：

$$\mathcal{L}_{\mathrm{emb}}(\mathrm{D}) = \frac{1}{\mathrm{D}} \sum_{(\mathrm{X},\mathrm{Y}) \in \mathrm{D}} \ell_{\mathrm{emb}}(\mathrm{X}, \mathrm{Y}), \tag{2.11}$$

其中 D 是机器翻译的训练集。

最后，神经机器翻译模型的总损失函数是翻译损失和正则项的加权和：

$$\mathcal{L} = \mathcal{L}_{\mathrm{mt}}(\mathrm{D}) + \alpha \mathcal{L}_{\mathrm{emb}}(\mathrm{D}), \tag{2.12}$$

这里 α 是控制语义相关正则化比重的超参。

语义相关正则化借鉴了 Word2vec 的思想，在方法实现上对 Skip-gram 进行了以下方面的拓展：（1）从单语场景拓展到双语场景，通过一个目标函数同时

增强词向量的单语相关性和对齐相关性；（2）将训练方式拓展到句级别的并行训练，让中心词在一次预测中预测其所有窗口词，训练标签是关于所有窗口词的平均分布（公式 2.10），使得语义相关正则化能够和神经机器翻译模型共同训练。(3) 方法中的输入层词向量和输出层词向量都使用神经机器翻译模型的词向量参数代替（公式 2.9），无需增加额外的参数。

2.5　实验

本章在 NIST 中文-英文数据集对语义相关正则化进行了实验，并把它和相关的词向量正则化方法进行了比较。

2.5.1　数据集

NIST 中文-英文数据集包含 125 万行平行语料，其中英文数据使用了 Moses 进行 tokenize，中文数据使用了斯坦福分词工具进行 tokenize。数据集的词表经过了 3.2 万次合并操作的 BPE 处理。本章使用 mt02、mt03、mt04、mt05、mt06、mt08 数据集作为测试集。

本章使用 NIST 中英数据集的原因有如下几点：（1）NIST 中英数据集大小适中，是机器翻译最常用的数据集之一，在该数据集上的实验结果具有较强的说服力；（2）中文和英文是字符系统不同的两种语言，具有较强的普适性，也是翻译需求比较大的翻译方向。(3) 中文和英文是作者掌握的语言，在进行语义分析时比较方便。

2.5.2　参数设置

为了公平地比较所有方法，本章在 Transformer (Vaswani 等, 2017) 模型上对所有方法进行了复现。本章的实验都是在开源工具 Fairseq (Ott 等, 2019) 的基础上进行的，使用了完全相同的实验配置和硬件设备。

本章的实验使用了 Transformer base 的实验配置，编码器和解码器都是 6 层，模型维度为 512 维，注意力机制使用了 8 个头（Head），前馈网络为 2048 维。所有模型都使用了 4000 个预热步（warm-up steps），初始学习率为 $7e^{-4}$，标签平滑比重为 0.1，Dropout 概率为 0.1，Adam 的参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$ 和 $\epsilon = 1e^{-9}$(Vaswani 等, 2017)。所有的实验都是在 4 块 GeForce RTX 3090 图像处理器进行的，批量训练的大小（Batch size）为 4096×8 个 token。在测试时，柱搜

索参数（Beam Size）为 4，长度惩罚（Length Penalty）为 0.6。

中英数据集中文和英文的词表大小分别为 40072 和 29408，模型一共训练了 35 轮。本章设置公式 2.12 中的超参数 $\alpha = 2$。

2.5.3 对比方法

为了对比，本章复现了五个相关性工作：

- 权重共享（Weight Tying, WT）(Press 等, 2017)。神经机器翻译模型的词向量参数分为三部分：源端输入词向量（Encoder Input Embeddings）、目标端输入词向量（Decoder Input Embeddings）、目标端输出词向量（Decoder Output Embeddings）（类别特征向量）。对于所有语言对，原则上可以共享目标端的两个词向量参数 (Decoder Weight Tying, Decoder-WT)；对于使用相同字符系统的语言对，可以将源端和目标端的词表进行合并，共享所有三个词向量参数（Three Way Weight Tying, TWWT）。

- 句级别对齐（Sentence Level Agreement, Sent-agree）(Yang 等, 2019)。作者将模型编码器输出的句级别向量和解码器输入层的句级别向量的 L_2 距离作为正则项来促进两端的信息一致。这里句级别向量是通过对句子中的词向量进行平均得到的。

- 直接链接（Direct Bridging）(Kuang 等, 2018)。作者以编码器和解码器的注意力权重作为对齐依据，拉近对齐的源端词和目标端词的距离。

- 球面词向量（Spherical Embedding）(Meng 等, 2019)。在球面词向量中，所有词向量的范数都被限制为 1，以解决数据不平衡造成的词向量范数差异过大的问题。

- 部分共享词向量（Shared-private Bilingual Word Embedding, Shared-private）(Liu 等, 2019)。作者利用词对齐工具获取词对齐关系，共享对齐词的词向量的部分维度，而将剩余未共享的维度作为独立学习的参数。例如对于对齐词“访问-visits”来说，两个词向量共享 90% 的维度，各自只有 10% 的维度是独立学习的。

2.5.4 实验结果

表 2.1 展示了所有词向量正则化方法在 NIST 中文-英文数据集上的实验结果。其中，除了句级别对齐（Sent-agree）方法之外，所有词向量正则化方法相比

表 2.1 在 NIST 中文-英文数据集上的实验结果

Table 2.1 Experimental results on NIST Zh-En task

	NIST Zh→En							
	mt02	mt03	mt04	mt06	mt06	mt08	AVG	Δ
Vanilla	45.31	44.13	45.82	43.9	43.72	34.9	42.96	–
Decoder-WT	45.82	45.77	46.07	44.43	44.27	34.63	43.50	+0.54
Sent-agree	45.32	44.06	45.88	43.83	43.81	34.74	42.94	-0.02
Direct Bridging	45.63	45.22	45.83	44.14	44.02	34.7	43.26	+0.30
Spherical	45.05	44.12	45.69	44.22	44.27	34.62	43.00	+0.04
Shared-private	46.41	45.73	46.7	45.04	44.77	35.71	**44.06**	**+1.10**
Our model	46.09	46.44	47.08	46.57	44.92	36.79	**44.65**	**+1.69**

基线模型（Vanilla）都取得了一定的提升，其中使用外部对齐工具的部分共享词向量（Shared-private）方法的效果最好，相比基线模型（Vanilla）取得了 1.10 个 BLEU 的提升。

语义相关正则化方法显著超过了所有的对比方法，相比基线模型（Vanilla）提升了 1.69 个 BLEU ，相比部分共享词向量（Shared-private）方法提升了 0.59 个 BLEU 。实验结果充分体现了语义相关正则化的有效性。

2.5.5　对齐相关性分析

本小节探究语义相关正则化对词向量参数语义特征的具体影响。本小节利用模型在人工对齐的测试集（NIST 05 测试集）上的对齐错误率（AER）分析了神经机器翻译模型的词向量参数和每一层隐状态的对齐性。所有对齐关系都提取自解码器向量对于编码器向量的注意力权重，即每个目标端词的源端对齐词通过注意力的最高权重来确定。由于模型每层隐状态的注意力有 8 个头，本章使用所有注意力头的平均权重作为对齐依据。

如表 2.2 所示，由于没有显式的促进语义相关性的训练目标，基线模型（Vanilla）的词向量具有较高的对齐错误率（93.24%），而语义相关正则化训练的词向量参数取得了 38.49% 的对齐错误率，具备非常强的对齐相关性，这说明本章的方法能够有效地增强词向量的语义相关性。

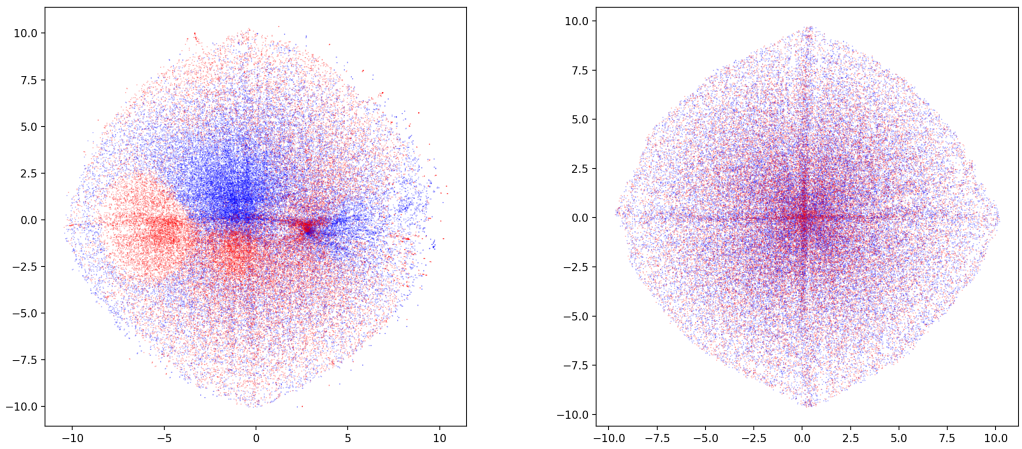
表 2.2 对齐错误率 (AER) 分析

Table 2.2 Analysis of AER

	NIST Zh→En						
	Embed	L1	L2	L3	L4	L5	L6
Vanilla	93.24	97.93	94.93	94.62	81.61	52.7	69.66
Our model	**38.49**	89.45	89.96	88.29	89.47	86.69	87.89

值得注意的是，基线模型（Vanilla）的对齐性在模型第 5 层达到了最优（52.7%），Li 等 (2019a) 将该现象归结于神经机器翻译模型会先通过前几层建立一定的对齐性，然后在对齐的基础上进一步学习复杂的高层表示信息。语义相关正则化方法虽然极大提高了词向量参数的对齐性，但是其高层隐状态的对齐错误率却稳定在 88% 左右，本章推测该现象是因为模型在词向量具备较强对齐性的情况下，不再需要前几层网络建立对齐关系，而是直接将每层网络用于高层表示信息的学习。

2.5.6 词向量的可视化



(a) **(b)**

图 2.3 词向量参数的可视化

Figure 2.3 Visualization of Word embeddings

如图 2.3 所示，本小节将神经机器翻译模型的词向量进行了 t-sne 降维和可

视化。其中蓝色为源端的词向量空间，红色为目标端的词向量空间，2.3 (a) 是基线模型的词向量空间，2.3 (b) 是使用语义相关正则化方法训练的词向量空间。可以看到，在基线模型的词向量空间中，源端词向量和目标端词向量分布在空间的不同位置，两者具有明显的分布边界；经过语义相关正则化训练后，词向量空间中源端词向量和目标端词向量较为均匀地融合到了一起，这和本章的目标相一致。

2.5.7　消融实验

表 2.3 在 NIST 中文-英文数据集上的消融实验

Table 2.3 Ablation Study on NIST zh-en task

	NIST Zh→En							
	mt02	mt03	mt04	mt06	mt06	mt08	AVG	Δ
基线模型	45.31	44.13	45.82	43.9	43.72	34.9	42.96	–
单语-源端	45.53	45.38	46.12	45.18	44.55	35.4	43.69	+0.73
单语-目标端	45.66	45.06	46.2	44.9	44.18	35.26	43.54	+0.58
单语-双端	45.84	45.81	46.54	45.61	44.54	35.8	44.02	+1.06
融合-源端	46.26	46.05	46.8	45.73	44.93	36.05	44.30	+1.34
融合-目标端	46.22	45.81	46.85	45.65	44.85	35.89	44.21	+1.25
融合-双端	46.09	46.44	47.08	46.57	44.92	36.79	**44.65**	**+1.69**

本小节对语义相关正则化中单语相关性的贡献和模型源端、目标端的贡献进行了消融实验。具体上，本小节分别对源端、目标端、双端都进行了单语的 Skip-gram 训练以衡量单语相关性对模型性能的贡献。对于语义相关正则化，本小节还分别以源端词作为中心词或目标端词作为中心词进行了非对称的训练，对比它和单语相关性的性能。

如表 2.3 所示，分别促进源端词向量的单语相关性或目标端的单语相关性都能够提升模型的性能，两端同时使用的效果更好，比基线模型提升 1.06 个 BLEU。非对称地使用语义相关正则化比仅提升某端的单语相关性效果更好，例如“融合-源端”的效果比“单语-源端”的效果更好。另外本章发现，对源端的词向量进行正则化比目标端更加有效，本章认为原因在于源端的词向量参数距离模型

的输出层比目标端更远，同时源端没有自回归语言模型任务的约束，因此词向量参数学习的程度更低，对正则化需求更高。

2.6 本章小结

模型参数的复杂度是模型过拟合的原因之一。对于神经机器翻译模型，词向量参数占模型总参数量的一半左右，具有较高的冗余度；同时词向量参数位于模型的两个终端，通过翻译损失的梯度更新参数，无法充分学习到语义特征。为此，本章提出语义相关正则化方法以提升神经机器翻译模型词向量参数的语义相关性。本章的方法完全基于双语语料的内在特征，适用于所有翻译方向且不依赖外部的知识指导。具体而言，机器翻译的一个训练样本由一个源端句子和一个目标端句子组成，其中隐含了单语言的词共现信息和双语言之间的词共现信息。基于这两种共现信息，本章提出一个自编码的训练目标来同时促进词向量的单语相关性和对齐相关性。实现上，本章借鉴 Skip-gram 的思想，将平行句对看作一个整体的序列，让其中的每个词预测序列中的其他词，从而使得频繁共现在平行句对的词学习到相近的词向量。本章的方法将 Skip-gram 从单语场景拓展到了双语场景，通过一个目标函数融合了对齐相关性和单语相关性两个方面，并将训练方式拓展到句级别的并行训练，使方法能够和神经机器翻译模型共同训练。在 NIST 中文-英文数据集上，本章的方法取得了 1.69 BLEU 的提升，显著超过了以往的词向量正则化方法。分析实验表明，语义相关正则化方法训练的词向量具有很强的语义相关性，两端词向量在空间的分布也更加均匀一致。

第 3 章　基于预测差异的扰动正则化方法

3.1　引言

神经机器翻译模型过拟合的表现之一是对噪声的鲁棒性很差，在输入中添加少量的噪声就会导致模型的性能出现大幅度的下降。近年来，在训练时在模型输入中添加扰动（即微小的噪声）来提升模型对噪声的鲁棒性、降低模型过拟合的扰动正则化方法（Perturbation Regularization）得到普遍的关注，并被成功地应用于机器翻译任务上 (Bengio 等, 2015; Wu 等, 2019; Sato 等, 2019a; Takase 等, 2021a)。扰动正则化方法的初衷是提升模型对噪声的鲁棒性，因为神经网络模型的泛化性能较差，在受到微小噪声的攻击后性能大幅度下降 (Bengio 等, 2015; Wu 等, 2019; Sato 等, 2019b; Takase 等, 2021b)。在扰动正则化方法中，模型的训练目标是最大化被扰动样本的似然（likelihood），这样可以改善模型在噪声干扰下的性能。在过去几年，多种类型的扰动被应用于神经机器翻译模型并展现出了良好的性能，包括词删除（Word-Dropout，WD）(Gal 等, 2016)、词替换（Word-Replacement，WR）(Bengio 等, 2015; Wu 等, 2019) 和对抗性扰动（Adversarial Perturbation, Adv）(Miyato 等, 2017; Sato 等, 2019b) 等。

已有的扰动正则化工作通常致力于寻找更有效的扰动形式和更合理的扰动策略，本章则从训练机制的角度入手，重新审视了当前“扰动-拟合”的训练机制。本章认为已有的训练机制忽略并加剧了模型对训练数据的欠拟合，限制了扰动正则化的性能。本章进一步提出预测差异正则化（Prediction Difference Regularization, PD-R ）来同时缓解模型的过拟合和欠拟合问题。本章的方法简单且有效，极大地提升了扰动正则化的性能。

具体而言，本章利用模型在噪声干扰前后对词级别标签的预测差异来分析词级别样本的拟合情况。通过定量的实验分析，本章发现，模型的输入被噪声干扰后，模型对相当一部分标签的预测概率不降反升，这说明模型对原始样本的拟合程度低于对被扰动样本的拟合程度，即模型对原始样本是“相对欠拟合”的。已有的扰动正则化工作忽略了这一点。本章根据实时的预测差异进一步将词级别标签划分为“相对欠拟合”集合和“相对过拟合”集合，并对词删除扰动正则化进行了选择性训练，即每次只让一个集合的标签去拟合被扰动的输入，而另一

个集合去拟合原始的输入。实验结果表明，让相对欠拟合的标签拟合被扰动的输入，会严重降低模型的性能，而相反做法的效果超过了已有扰动正则化方法无差别的做法。以上分析和实验证明，对相对欠拟合样本进行加噪训练会降低模型的性能，已有扰动正则化方法对所有样本进行无差别加噪训练的做法不是最优的。

本章进一步提出将预测差异作为正则项来改进扰动正则化方法。这里的预测差异指的是由输入的扰动引起的预测分布的差异。由于预测差异反映了模型对词级别样本的相对过拟合和相对欠拟合的程度，约束预测差异来同时减轻过拟合和欠拟合是一个自然的解决方案。通过组合交叉熵损失和预测差异正则项，模型在拟合训练数据的同时能够对过拟合和欠拟合保持一定的控制。

本章在最简单的词删除扰动上应用了预测差异正则化，并在三个常用的 WMT 数据集上进行了实验。这三个数据集囊括了小型、中型和大型数据集，实验结果具有较强的说服力。相比已有的扰动正则化方法，本章的方法取得了巨大的提升。在 WMT16 En-De 上，本章的方法相比基线模型提升了 1.80 个 SacreBLEU，相比已有的扰动正则化方法提升了 1.12 个 SacreBLEU。

3.2 相关工作

本章利用预测差异分析和改进了扰动正则化方法，相关工作主要涉及三个方面：在输入层添加扰动的工作、分析扰动的影响的工作以及与预测差异相关的工作。

3.2.1 与输入层扰动相关的工作

近年来，研究者们在自然语言处理领域提出了一些扰动正则化方法，这些方法在模型的输入中添加噪声以提升模型对噪声的鲁棒性。Gal 等 (2016) 提出对循环神经网络语言模型施加词删除噪声，Bengio 等 (2015) 和 Wu 等 (2019) 提出对神经机器翻译模型施加词替换噪声，Miyato 等 (2017) 和 Sato 等 (2019b) 分别提出对文本分类模型和神经机器翻译模型施加对抗性扰动，以上工作显著增强了神经网络模型的泛化性和对噪声的鲁棒性。

除了扰动正则化方法之外，一些工作还将不确定性引入了子词（Subword）划分阶段以提升模型的泛化性能和对子词划分的鲁棒性。Kudo (2018b) 提出为子词划分提供多个候选并在训练过程中随机采样，Provilkov 等 (2020) 提出在对子词划分施加 Dropout，Park 等 (2020b) 提出对抗性的子词划分。对于字符级别

（Character-level）的自然语言处理任务，一些工作提出对模型输入施加字符级别的噪声，包括删除（deletion）、插入（insertion）、替换（substitution）、置换（swap）(Belinkov 等, 2018; Karpukhin 等, 2019) 以及对抗性字符替换 (Ebrahimi 等, 2018) 等。另外，自然语言处理中的 Mixup 技术使用对样本进行了混合，也可以看作是一种扰动，该技术被应用于数据增强或者多样性生成 (Guo 等, 2020a; Li 等, 2021; Fang 等, 2022)。

3.2.2　与扰动的影响相关的工作

在以往的工作中，扰动通常被视认为是一种负面因素，一些工作发现细微的噪声会导致神经网络模型产生错误的结果。(Szegedy 等, 2014) 发现图像中不可察觉的扰动会导致图像分类模型分类错误，Liang 等 (2018) 发现对文本输入添加扰动会“欺骗”文本分类模型，Belinkov 等 (2018) 发现对源端句子添加微小的噪声会导致机器翻译模型生成错误的译文，模型的性能会随着对输入的修改比例增加而单调下降，这和本章的实验结果是一致的。Khayrallah 等 (2018) 和Briakou 等 (2021) 发现把机器翻译的训练样本部分替换为噪声或者语义不同的文本会降低模型训练的性能。

3.2.3　与预测差异相关的工作

输入中的变动带来的预测差异体现了模型输入和输出的关系，因此通常被用于模型的行为分析。Zintgraf 等 (2017) 提出使用预测差异分析图像的局部区域对图像分类模型决策的影响，Li 等 (2019a) 通过每次删除源端的一个词并计算目标端标签的预测概率变化来分析神经机器翻译模型的词对齐关系。另外，Guo 等 (2019) 发现可以通过预测差异来探测对抗性样本，Liang 等 (2021) 提出将子模型之间的预测差异作为正则项来约束模型参数学习的自由度。

3.3　研究背景

扰动正则化方法在训练过程中对模型输入添加扰动（细微的噪声）来提升模型对噪声的鲁棒性。本节将介绍扰动正则化的主要类型以及对应的目标函数。

3.3.1　常用扰动类型

本节将介绍三种常见的扰动类型：括词删除（Word-Dropout，WD）、词替换（Word-Replacement，WR）和对抗性扰动（Adversarial Perturbation, Adv）。

- 词删除和词替换。对输入施加扰动的最简单的方法是直接改变原始输入序列中的词，其中词删除和词替换是最常用的形式。由此产生的被扰动序列 $\hat{x}$ 是从原始输入序列和噪声序列中采样的结果：

$$\hat{x_i} = \begin{cases} x_i, & \text{with probability 1 - } \alpha, \\ x_i^p, & \text{with probability } \alpha, \end{cases} \tag{3.1}$$

这里 $0 < \alpha < 1$ 是伯努利采样的超参，x_i^p 是噪声序列 x^p 的第 i 个词。需要注意的是，对于词删除 (Gal 等, 2016) 来说，噪声序列 x^p 由零向量组成，而对于词替换 (Bengio 等, 2015; Wu 等, 2019) 来说，噪声序列 x^p 是通过平均分布或者某个特定的分布从词表中随机采样得到的。

- 对抗性扰动。对抗性扰动正则化认为能够最大化模型损失的单位扰动向量是最有效的扰动。如同Miyato 等 (2017) 和 Sato 等 (2019b) 的定义，被扰动的词向量可以通过如下计算得到：

$$\hat{\mathbf{e}}_i = \mathbf{e}_i + \kappa \hat{\mathbf{r}}_i, \tag{3.2}$$

其中 $\mathbf{e}_i$ 是 x_i 原始的词向量，κ 是控制扰动向量大小的超参，$\hat{\mathbf{r}}_i$ 是通过对模型损失函数进行梯度下降获取的最强的单位扰动向量 (Goodfellow 等, 2015)：

$$\hat{\mathbf{r}}_i = \frac{\mathbf{g}_i}{||\mathbf{g}_i||_2}, \quad \mathbf{g}_i = \nabla_{\mathbf{e}_i} \mathcal{L}(\mathrm{D}, \theta), \tag{3.3}$$

这里 $\mathbf{g}_i$ 是损失函数相对于词向量 $\mathbf{e}_i$ 的梯度。

在大部分场景下，我们不仅可以对编码器的输入进行扰动，还可以以同样的方式对解码器的输入进行扰动。需要注意的是，由于 Scheduled Sampling 面向机器翻译的曝光偏差问题，该方法的词替换扰动仅应用于解码器端。

3.3.2 损失函数

- 词删除和词替换。对于词删除和词替换来说，模型的训练目标是拟合被扰动的训练样本：

$$\mathcal{L} = \mathcal{L}(\hat{\mathrm{D}}, \theta) = -\frac{1}{\mathrm{D}} \sum_{(\hat{\mathrm{X}}, \mathrm{Y}) \in \hat{\mathrm{D}}} \ell(\hat{\mathrm{X}}, \mathrm{Y}, \theta), \tag{3.4}$$

这里 $\hat{\mathrm{D}}$ 被扰动后的数据集。

- 对抗性扰动。对于对抗性扰动来说，模型在训练时需要两次前向传播（Forward Propagation）和两次反向传播（Backward Propagation）。第一次前向传播计

算出模型对原始样本的损失，第一次反向传播计算了模型相对于原始样本的梯度，并计算出对抗性扰动；第二次前向传播使用了第一次反向传播计算出的对抗性扰动，计算出模型对被扰动样本的损失，第二次反向传播计算了模型损失相对于被扰动样本的梯度。最终，模型通过两次梯度回传，同时拟合原始样本和被扰动样本：

$$\mathcal{L} = \mathcal{L}(\mathrm{D}, \theta) + \lambda \mathrm{L}(\hat{\mathrm{D}}, \theta), \tag{3.5}$$

这里 λ 是控制扰动正则化的超参。对于对抗性扰动，扰动被应用在词嵌入层（Embedding Layer）而不是输入层（Input Layer）。

3.4　预测差异正则化

本节首先利用输入层噪声引起的预测差异分析了神经机器翻译模型对词级别样本的拟合情况，其次提出将预测差异作为正则项来同时减轻模型的过拟合和欠拟合。

3.4.1　基于预测差异的样本拟合分析

已有工作通常认为模型对标签的预测能力在受到噪声干扰后会下降，因此扰动正则化方法被提出以增强模型对噪声的鲁棒性。然而噪声干扰下模型对标签的预测概率并不一定会下降，这主要有以下三个原因：

- 模型的复杂性。神经网络模型拥有巨大数量的参数，可解释性较差，模型的行为不一定符合人类逻辑。
- 扰动的复杂性。扰动的生成通常是随机的、不可控的，在这种情况下，扰动不仅可能是噪声，还有可能是近义词或者同义词，这种情况下，模型的预测可能不会受到负面影响。
- Dropout 的影响。在实验中，模型的训练一般使用 Dropout 技术以减轻模型过拟合，这样噪声干扰前后的两轮计算使用了两个不同的子模型（Sub Model），模型对标签的预测概率受到随机子模型的影响，因此不一定会下降。

为了验证噪声干扰对模型预测能力的影响，本节对词级别标签的预测差异进行了分析实验，这里预测差异被定义为模型在噪声干扰前后对词级别标签预

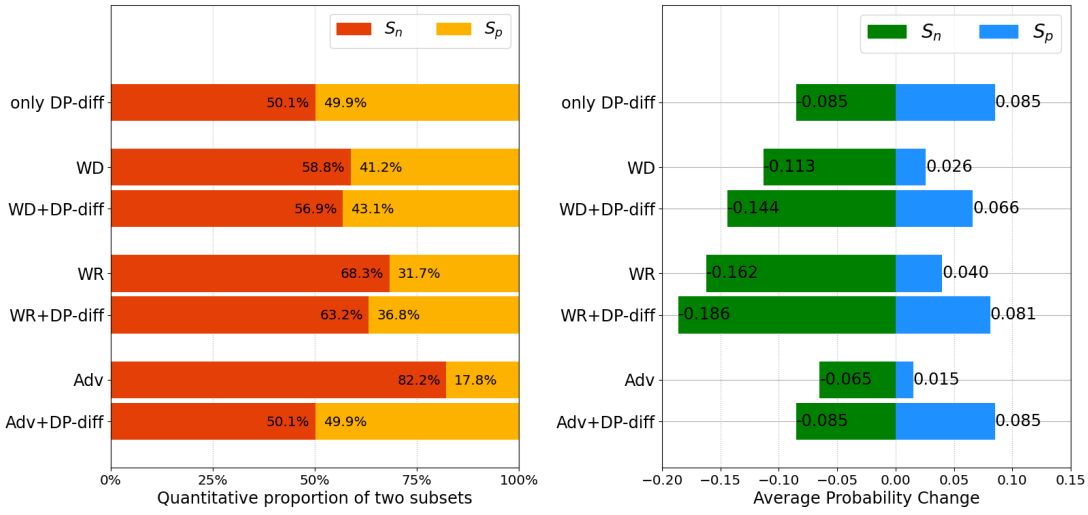
测概率的变化 (Gu 等, 2019; Li 等, 2019a)：

$$\begin{aligned}\Delta p(y_j) = & p(y_j|\hat{y}_{<j}, \hat{\mathrm{X}}, \theta) \\ & - p(y_j|y_{<j}, \mathrm{X}, \theta).\end{aligned} \tag{3.6}$$

在测试集上，本章使用各种类型的扰动对模型的输入进行了干扰，并根据预测差异把所有的词级别标签划分为两个子集合：预测概率上升的标签组成的正反馈集合 $\mathbf{S}_p$ 和预测概率下降的标签组成的负反馈集合 $\mathbf{S}_n$。本章定量统计了两个集合的数量比和平均概率变化以分析模型对不同类型扰动的反馈。

由于预测差异还受到参数 Dropout 的影响，本章通过控制计算预测差异的两轮前向传播的 Dropout 掩码（Mask）相同或者不同，来具体分析 Dropout 对预测差异的影响。如果两轮的 Dropout 掩码不同，原始输入的前向传播和被扰动输入的前向传播分别使用了两个不同的子模型。如果两轮 Dropout 掩码相同，预测差异就排除了子模型差异的影响。

实验上，本章使用了在 WMT16 En-De 数据集上训练好的模型，在测试集上进行了扰动实验，该测试集是 WMT16 到 WMT20 5 个测试集的整合。本章的分析实验包含三种扰动类型：词删除、词替换和对抗性扰动。其中词删除和词替换的概率是 0.05，所有的扰动都同时应用于编码器端和解码器端。



(a) **(b)**

图 3.1 不同类型的扰动对词级别标签预测的影响

Figure 3.1 Influence of different perturbations on token-level label prediction

图3.1（a) 和图3.1（b) 分别展示了两个子集的数量比和平均预测概率变化。

可以看到，扰动对词级别标签预测的负面影响是主要的，这个结论对各种类型的扰动都成立。但值得注意的是，扰动的正面影响也是不容忽视的，对于词删除和词替换，正反馈的标签比例甚至可以达到 30%-40%。如果加上参数 Dropout 的影响，扰动的正面影响还会继续增大。

本章将预测差异归结于模型对词级别样本的“相对过拟合”和“相对欠拟合”。由于扰动是比重非常小的噪声，因此被扰动的样本仍然可以看作是好的样本。在这个前提下，如果一个词级别的预测概率下降，说明模型对原始样本是“相对过拟合”的，而预测概率上升则说明模型对原始样本是“相对欠拟合”的。受到参数 Dropout 的影响，这些拟合问题还反映了子模型对于原始样本的拟合偏好，这也是需要避免的。

上文提到，已有扰动正则化方法的前提是扰动对模型预测只有负面影响，但是本章的分析表明扰动还有可能提升模型对标签的预测能力。这意味着在某些情况下，让模型去拟合被扰动样本是不必要的。

表 3.1 对词删除扰动正则化进行选择性训练的实验结果

Table 3.1 BLEU (Papineni 等, 2002) for selective training of word-dropout perturbation

	En→Ro	En→De
Transformer	33.4	32.55
only $\mathbf{S}_p$	31.59	30.16
only $\mathbf{S}_n$	**34.57**	**33.73**
$\mathbf{S}_p$ and $\mathbf{S}_n$	34.53	33.20

本章进一步对词删除扰动正则化进行了选择性训练，即每次只让一个集合的标签去拟合被扰动的输入，而另一个集合去拟合原始的输入。如图3.1所示，只对欠拟合子集合 $\mathbf{S}_p$ 进行加噪训练，会降低模型的性能，而只对过拟合子集合 $\mathbf{S}_n$ 进行加噪训练的实验效果超过了已有扰动正则化方法无差别训练的做法。这说明已有的扰动正则化方法受制于模型对训练数据的欠拟合。

3.4.2 预测差异正则化

由于预测差异反映了模型对原始样本的“相对过拟合”和“相对欠拟合”，而这两种情况都是我们希望避免的，因此本章提出直接把预测差异作为正则项:

$$\ell_{\text{PD-R}}(\mathrm{X}, \mathrm{Y}, \theta) = \mathcal{R}[\mathrm{P}(*|\mathrm{X}, \mathrm{Y}_{<}, \theta'), \mathrm{P}(*|\hat{\mathrm{X}}, \hat{\mathrm{Y}}_{<}, \theta'')], \tag{3.7}$$

其中 $\mathcal{R}[\cdot]$ 为两轮预测分布的距离，(X, Y) 是数据集 D 中的双语句对，$*$ 表示目标端的所有标签，$\mathrm{P}(*|\mathrm{X}, \mathrm{Y}_{<}, \theta')$ 是条件于源端输入 X、目标端输入 $\mathrm{Y}_{<}$ 和子模型 θ' 的所有目标端标签的预测分布，$\mathrm{P}(*|\hat{\mathrm{X}}, \hat{\mathrm{Y}}_{<}, \theta'')$ 是条件于源端被扰动输入 $\hat{\mathrm{X}}$、目标端被扰动输入 $\hat{\mathrm{Y}}_{<}$ 和子模型 θ'' 的所有目标端标签的预测分布。正则化的损失是对所有数据集样本的平均:

$$\mathcal{L}_{\text{PD-R}}(\mathrm{D}, \theta) = \frac{1}{\mathrm{D}} \sum_{(\mathrm{X},\mathrm{Y})\in\mathrm{D}} \ell_{\text{PD-R}}(\mathrm{X}, \mathrm{Y}, \theta). \tag{3.8}$$

训练模型的总损失函数是交叉熵和正则项的加权和:

$$\mathcal{L} = \mathcal{L}(\mathrm{D}, \theta) + \gamma \mathcal{L}_{\text{PD-R}}(\mathrm{D}, \theta), \tag{3.9}$$

这里 γ 是控制正则化比重的超参。

在实验中，本章将 PD-R 应用在最简单的词删除扰动上，词删除的比例是 0.05，即公式 (3.1) 中 $\alpha = 0.05$。同时，本章将公式 (3.9) 中的超参设置为 $\gamma = 1.0$，没有进行大规模的超参数搜索。公式 (3.7) 中的 $\mathcal{R}[\cdot]$ 具体实现为 L_1 距离，其效果在实验中比 KL 距离稍好一些。

3.5 实验

本章在三个常用的 WMT 数据集对预测差异正则化（PD-R）进行了实验，并把它和相关的对比方法进行了比较。

3.5.1 数据集

为了充分验证预测差异正则化（PD-R）的效果，本章在三个常用的 WMT 数据集进行了实验，这三个数据集分别为小型的 WMT16 英文-罗马尼亚文数据集、中型的 WMT16 英文-德文数据集和大型的 WMT17 中文-英文数据集。

- WMT16 英文-罗马尼亚文数据集包含大约 60 万行平行语料，由 Moses 工具 (Koehn 等, 2007b) 进行 tokenize，词表进行了 4 万次合并操作的 BPE (Sennrich

等, 2016b) 处理。本章分别使用 news-dev 2016 和 news-test 2016 作为验证集和测试集。

- WMT16 英文-德文数据集包含大约 450 万行平行语料，词表进行了 3 万次合并操作的 BPE 处理。测试时，本章对最后 5 个模型（Epoch）进行了参数平均的操作，并使用了从 WMT 2016 到 WMT 2020 的 5 个测试集的平均 BLEU 来衡量模型的效果。
- WMT17 中文-英文数据集包含超过 2000 万行平行语料。其中英文数据使用了 Moses 工具进行 tokenize，中文数据使用了斯坦福分词工具进行 tokenize。数据集的词表进行了 3.2 万次合并操作的 BPE 处理。本章分别使用 newsdev2016 和 newsdev2017 作为验证集和测试集。

3.5.2　参数设置

为了公平地比较所有方法，本章在 Transformer (Vaswani 等, 2017) 模型上对所有方法进行了复现。本章的实验都是在开源工具 Fairseq (Ott 等, 2019) 的基础上进行的，使用了完全相同的实验配置和硬件设备。

本章的实验使用了 Transformer base 的实验配置，编码器和解码器都是 6 层，模型维度为 512 维，注意力机制使用了 8 个头（Head），前馈网络为 2048 维。所有模型都使用了 4000 个预热步（warm-up steps），初始学习率为 $7e^{-4}$，标签平滑比重为 0.1，Adam 的参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$ 和 $\epsilon = 1e^{-9}$(Vaswani 等, 2017)。对于小型的英罗数据集，Dropout 概率为 0.2，对于中大型的英德和中英数据集，Dropout 概率为 0.1。所有的实验都是在 4 块 GeForce RTX 3090 图像处理器进行的，批量训练的大小（Batch size）为 4096×8 个 token。在测试时，柱搜索参数（Beam Size）为 4，长度惩罚（Length Penalty）为 0.6。

对于英罗和英德数据集，本章共享了源端和目标端词表，词表大小均为 32768，在训练时使用了 TWWT (Press 等, 2017) 的设置，模型一共训练了 50 轮。对于中英数据集，中文和英文的词表大小分别为 44000 和 33000，模型一共训练了 30 万步。

3.5.3　对比方法

为了对比，本章复现了四种具有代表性的扰动正则化方法以及最近发表的 R-Drop 工作。

- 词删除（Word-Drop）。词删除 (Gal 等, 2016) 将输入序列的词向量随机替换为零向量，公式 (3.1) 中 $\alpha = 0.05$。
- 词替换（SSE-SE）。SSE-SE (Wu 等, 2019) 是一种词替换方法，它将模型的输入词以一定概率随机替换为词表中的其它词，公式 (3.1) 中 $\alpha = 0.01$，扰动序列是以平均分布采样得到的。
- Scheduled Sampling。Scheduled Sampling (Bengio 等, 2015) 是一种仅应用于目标端的词替换方法，训练时输入词被随机替换为模型的预测词，替换概率 α 遵循课程学习的策略：

$$\alpha_i = \frac{k}{k + \exp(i/k)}, \tag{3.10}$$

这里 i 为训练步数，k 是依赖于模型收敛的超参，模型的实现是并行的 (Mihaylova 等, 2019; Duckworth 等, 2019)。对于英罗、英德和中英数据集，本章分别设置超参 $k = (4590, 29350, 36150)$，这里超参 k 的设置是为了保证模型的训练结束时 α_i 衰减为 0.9。

- 对抗性扰动（AdvT）。对于对抗性扰动正则化，本章参考Sato 等 (2019b) 设置公式 (3.2) 中 $\kappa = 1$，设置公式 (3.5) 中 $\lambda = 1$。
- R-Drop。R-Drop (Liang 等, 2021) 是一个在实现上和预测差异正则化非常类似的方法，但是该方法的目的是减少子模型的差异，而本章方法的目的是减少输入层扰动带来的预测差异以减少不恰当的拟合现象。由于子模型的差异也是预测差异来源的一部分，因此 R-Drop 可以看作是本章方法的一个部分。

3.5.4 实验结果

本章使用了 SacreBLEU (Post, 2018) 对译文进行评测，各模型的实验结果如表 3.2 和表 3.3 所示。本章分别在源端词删除、目标端词删除和两端词删除三种配置上使用了预测差异正则化（PD-R）。除 Scheduled Sampling 以外，所有扰动方法都同时应用于源端和目标端。同时，本章还对词删除扰动正则化进行了选择性训练（Selective Training, ST, only $\mathbf{S}_n$）以便与 PD-R 和普通的扰动正则化方法进行对比。

实验结果表明已有的扰动正则化方法之间的效果差异很小，这与Takase 等 (2021b) 的结果是一致的。R-Drop 和选择性训练（ST）的效果都显著超过了已有的扰动正则化方法。本章的方法 PD-R 在所有数据集上都超过了已有的扰动正则化方法和选择性训练（ST），并且在中小型数据集英罗和英德上显著超过了 R-Drop。

表 3.2 在 WMT16 英文-罗马尼亚文和 WMT17 中文-英文数据集上的实验结果

Table 3.2 Experimental results on WMT16 En-Ro and WMT17 Zh-En tasks

	WMT2016 En→Ro		WMT2017 Zn→En	
	2016	Δ	2017	Δ
Transformer (Vaswani 等, 2017)	33.16	–	23.98	–
Word-Drop	34.13	+ 0.97	24.20	+ 0.22
SSE-SE (Wu 等, 2019)	33.75	+ 0.59	24.14	+ 0.16
Scheduled Sampling (Bengio 等, 2015)	33.62	+ 0.46	23.74	– 0.24
AdvT (Sato 等, 2019b)	33.65	+ 0.49	24.17	+ 0.19
R-Drop (Liang 等, 2021)	34.14	+ 0.96	**25.08**	+ **1.10**
Word-Drop + ST	34.24	+ 1.08	24.37	+ 0.39
Word-Drop (enc) + PD-R	34.22	+ 1.06	**24.98**	+ **1.00**
Word-Drop (dec) + PD-R	34.57	+ 1.41	24.76	+ 0.78
Word-Drop (both) + PD-R	**34.93**	+ **1.77**	24.86	+ 0.88

表 3.3 在 WMT16 英文-德文数据集上的实验结果

Table 3.3 Experimental results on WMT16 En-De task

	WMT2016 En→De					
	2016	2017	2018	2019	2020	AVG
Transformer	33.81	27.75	40.56	36.39	21.95	32.09
Word-Drop	34.14	28.00	41.07	38.04	22.62	32.77(+0.68)
SSE-SE (Wu 等, 2019)	33.90	27.95	41.23	36.93	22.66	32.53(+0.44)
Scheduled Sampling	33.96	28.12	41.13	37.39	22.58	32.63(+0.54)
AdvT (Sato 等, 2019b)	34.25	27.91	41.31	37.05	23.00	32.70(+0.61)
R-Drop (Liang 等, 2021)	35.32	27.66	41.72	38.26	22.84	33.16(+1.07)
Word-Drop+ST	34.85	28.23	42.10	38.13	22.74	33.21(+1.12)
Word-Drop(enc)+PD-R	35.30	28.28	**42.92**	**39.09**	**23.88**	**33.89**(+**1.80**)
Word-Drop(dec)+PD-R	**35.39**	**28.34**	42.14	38.51	23.68	33.61(+1.52)
Word-Drop(both)+PD-R	35.17	28.23	42.20	38.74	23.61	33.59(+1.50)

在 WMT16 英德数据集上，PD-R 相比基线模型取得了 1.80 个 SacreBLEU 的提升，相比词删除扰动正则化取得了 1.12 个 SacreBLEU 的提升，相比 R-Drop 取得了 0.73 个 SacreBLEU 的提升。

在大型数据集 WMT17 中英上，PD-R 的效果没有超过 R-Drop。本章认为该现象的原因在于大型数据集的训练样本相对充足，因此数据层面的正则化效果不够明显，而模型参数层面的正则化能够保持较好的效果。

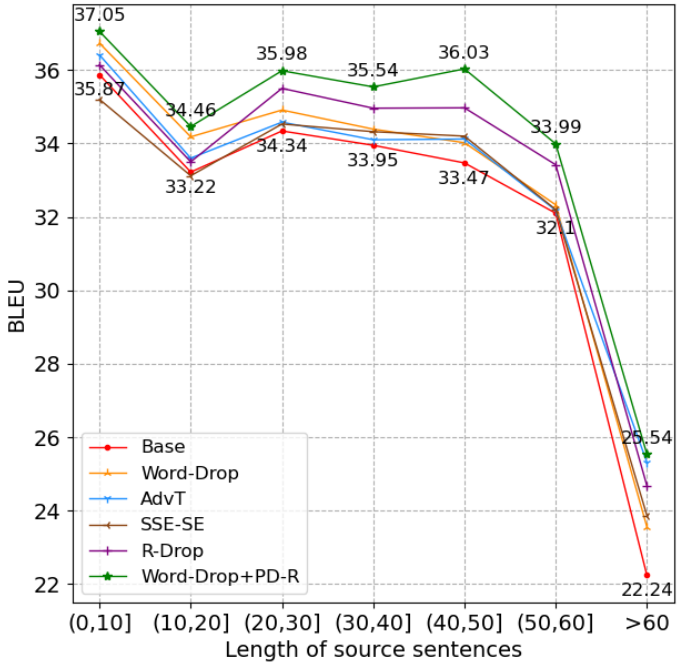


图 3.2 模型在不同长度的句子上的性能

Figure 3.2 Performance of models on sentences of different lengths

3.5.5 长句实验

长句通常包含更加复杂、少见的词组，长句翻译的曝光偏差也更加严重 (Ranzato 等, 2016; Zhang 等, 2019b)，因此在长句的性能体现了模型对于未知输入的鲁棒性。

本节在 WMT 英德数据集上进行了实验。本节使用了由 5 个测试集合并的测试集，并把测试集按照句长划分为 7 个子集。如图 3.2 所示，PD-R 在所有句长上的表现均超过了普通的扰动正则化方法，并且随着句长增加，性能提升也更大，这说明 PD-R 能够更好地处理长句中的未知输入。

3.5.6　噪声鲁棒性测试

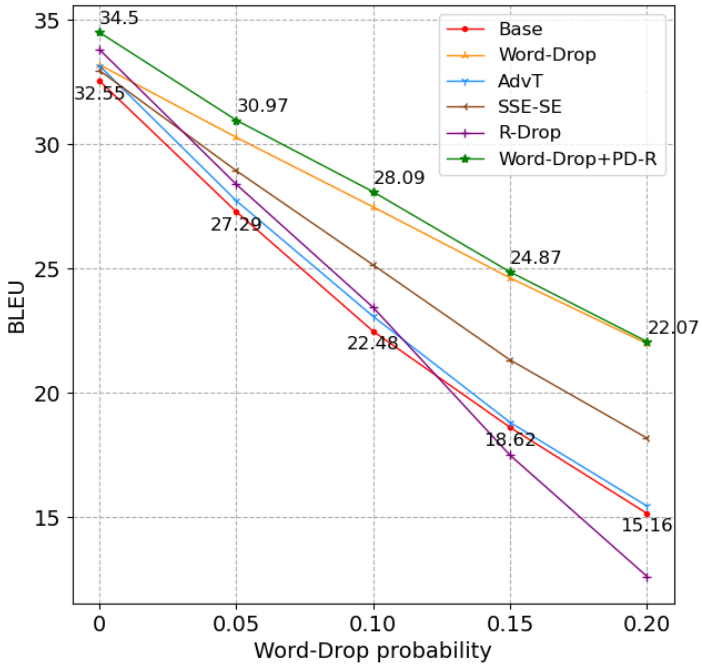


图 3.3 模型对词删除扰动的鲁棒性

Figure 3.3 Robustness of models to word-dropout perturbation

为了更好地衡量模型对扰动的鲁棒性，借鉴Michel 等 (2018) 和Moradi 等 (2021)，本节对各模型进行了噪声攻击实验。实验上，本节使用了源端的词删除扰动对模型进行攻击。如图 3.3 所示，PD-R 和词删除扰动正则化比基线模型对噪声的鲁棒性更强，这表明本章的加噪训练确实提升了模型对噪声的鲁棒性。

需要注意的是，本文在其他类型的噪声攻击实验表明，一个模型只有在训练时对某种噪声进行正则化训练之后才会对该种噪声攻击体现出更强的鲁棒性，因此与其它类型的扰动正则化方法的对比不是本分析实验的重点。

3.5.7　消融实验

消融实验的结果如表 3.4 所示，实验表明预测差异正则化对相对过拟合样本 $\mathbf{S}_n$ 和相对欠拟合样本 $\mathbf{S}_p$ 都有效，甚至对 $\mathbf{S}_p$ 的效果更好。这表明本章的方法可以较好地处理过拟合和欠拟合。

由于子模型的拟合偏好也是过拟合和欠拟合的来源之一，为了区分 Dropout 和词删除扰动各自对模型性能的贡献，本节分别设置两轮模型计算的 Dropout 掩

表 3.4 预测差异正则化方法的消融实验

Table 3.4 Ablation study for PD-R

	En→Ro	En→De
Transformer	33.40	32.55
only $\mathbf{S}_p$	34.83	34.22
only $\mathbf{S}_n$	34.59	34.16
$\mathbf{S}_p$ and $\mathbf{S}_n$	35.11	34.50
only DP-diff	**34.70**	**34.10**
WD(enc) w/o DP-diff	33.97	33.97
WD(dec) w/o DP-diff	34.15	33.15
WD(both) w/o DP-diff	**34.40**	**34.02**
WD(enc) w/ DP-diff	34.63	**34.50**
WD(dec) w/ DP-diff	34.92	34.19
WD(both) w/ DP-diff	**35.11**	34.22

码不同或相同来区分两者的贡献。同时本节分别对源端和目标端进行了扰动实验以区分源端扰动和目标端扰动的贡献。实验结果表明参数 Dropout 和词删除扰动都对预测差异正则化的性能提升产生了重要贡献，而两者一起使用的效果最好。至于源端和目标端的差别，在小型数据集上，目标端的扰动正则化比源端更加有效，但随着数据集的增加，源端的扰动正则化变得更加重要，这和主实验的结果是一致的。

3.6 本章小结

神经机器翻译模型对噪声的鲁棒性很差，在输入中添加少量的噪声就会导致模型的性能出现大幅度的下降。为此，研究者们提出扰动正则化方法，通过在训练样本中添加噪声以提升模型对噪声的鲁棒性。扰动正则化方法的前提假设是模型对原始样本的拟合程度总是高于对被扰动样本的拟合程度，但本章发现该前提并不总是成立。本章利用模型在输入受到扰动前后对目标词的预测差异分析了词级别样本的拟合情况，发现模型对相当一部分样本是相对欠拟合的，对这部分样本进行扰动正则化训练会损害模型的性能。为了同时缓解模型的过拟

合和欠拟合问题，本章提出预测差异正则化（PD-R），将输入层扰动引起的预测差异作为正则项训练神经机器翻译模型。在 WMT16 英德数据集上，预测差异正则化取得了 1.80 个 SacreBLEU 的提升，远远超过了已有的扰动正则化方法。分析实验表明，本章的方法显著提升了模型的噪声鲁棒性和泛化性。

第 4 章　基于条件变分自编码器的标签平滑方法

4.1　引言

由于训练数据的相对稀缺和使用硬标签（Hard Label）作为拟合目标的训练方式，神经机器翻译模型容易过拟合训练集中出现的标签而忽略其他可能正确的标签。为了缓解神经网络模型对硬标签过度自信（Over Confident）的问题，Szegedy 等 (2016) 首先提出使用平均分布对硬标签进行平滑，即使用硬标签与平均分布的加权作为拟合目标训练模型。该方法简单有效，被广泛应用于包括神经机器翻译的各种分类任务。然而，该方法将硬标签的概率平均分配给其它所有类别，认为所有类别的可能性是相等的，这个先验假设显然是不合理的。例如对于语言模型来说，给定前缀“美国总统访问”，硬标签为“中国”，在一个较好的平滑标签中，“日本”、“韩国”、“英国”等国家词汇的概率应该明显高于“苹果”、“天气”、“袭击”等不相关的词汇，而不是无差别地赋予所有词相同的概率。

另一种有效的标签平滑方法是知识蒸馏 (Hinton 等, 2015)。知识蒸馏起初被用于压缩模型，即使用训练好的大模型蒸馏小模型，从而在性能和效率之间取得平衡。近年来，一些工作 (Yuan 等, 2020; Zhang 等, 2020) 将知识蒸馏与标签平滑联系起来，认为知识蒸馏实际上是一种优秀的标签平滑方法。作为知识蒸馏的一种，自蒸馏 (Furlanello 等, 2018) 是使用训练好的相同结构、容量的模型来蒸馏重新初始化的模型，该方法能够明显提升模型的性能，是一种简单有效的标签平滑方法。但是，自蒸馏具有训练代价大、标签生成器过于臃肿的缺点，而已有的其它标签平滑方法性能都比较有限。本章致力于寻找一种高性能、高效率的在线标签平滑方法。

本章提出使用条件变分自编码器进行标签平滑。条件变分自编码器（Conditional Variational Auto Encoder, CVAE）首先被提出用于给定条件下的图像生成 (Sohn 等, 2015)，它能够将给定条件下的标签编码到一个和先验分布相近的后验分布空间并重构出标签，因此具有生成新标签的能力。本章希望利用条件变分自编码器建模给定源端和目标端输入的条件下词级别标签的隐变量分布，并从该分布采样出新的标签用于神经机器翻译模型的标签平滑。

实现上，本章在神经机器翻译模型上添加了一个轻量的 CVAE 模块，该模

块与翻译模型同时训练，并实时生成平滑标签。本章在 WMT16 英罗和 NIST 中英数据集上对该模型进行了实验。本章的方法显著超过了已有的在线标签平滑方法。相比离线的自蒸馏方法，本章的方法只需要 1.2 倍左右的训练时间；相比平均分布的标签平滑，本章的方法取得了 1.2 BLEU 的提升。

4.2 相关工作

本章基于条件变分自编码器建立了一个标签平滑模型，主要涉及标签平滑和变分自编码器的应用相关的工作。

4.2.1 标签平滑相关工作

Szegedy 等 (2016) 最早提出使用平均分布平滑硬标签以减轻模型对训练标签过度自信的问题，该方法被广泛应用于各种分类任务，但所有类别概率相同的假设显然是不符合实际场景的。针对该问题，研究者们陆续提出了一些改进方法。Luo 等 (2021) 认为在没有根据的情况下把硬标签的概率分给其他标签是不合理的，为此提出将硬标签的概率分给一个额外的虚假标签；Liu 等 (2021) 认为应该按照与硬标签的相似度将概率分配给其它标签，提出基于类别相似度的标签平滑方法并应用于图像分类任务；Zhang 等 (2021) 认为模型在训练过程中产生的标签可以用于自身的标签平滑，提出将每轮训练中模型产生的属于同一标签的预测分布的平均作为该标签在下轮训练的平滑标签，该方法在图像分类任务取得显著效果。另外，一些工作 (Yuan 等, 2020; Zhang 等, 2020) 认为知识蒸馏是一种标签平滑方法，其之所以有效是因为老师模型为学生模型提供了较好的平滑标签。除了以上工作，Xie 等 (2016) 提出随机干扰一些样本的标签以减轻模型对已有标签的过拟合；Reed 等 (2015) 提出使用模型自身预测的分布或模型自身预测的类别进行标签平滑；Dubey 等 (2018) 提出在类别映射中添加成对混淆（Pairwise Confusion, PC）以减轻模型过拟合；Li 等 (2020) 提出使用两个神经网络将图像和标签映射到隐空间并使用它们之间的随机距离对模型进行正则化。

4.2.2 变分自编码器相关工作

变分自编码器被广泛应用于图像样本的生成，在自然语言处理领域的应用相对较少。Bowman 等 (2015) 将变分自编码器引入语言模型来帮助循环神经网络建立句级别全局信息，该模型比常规的自然语言模型的性能稍差，但是具备更

强的文本生成能力；Miao 等 (2016b) 提出使用变分自编码器建立上下文信息并在文档建模和有监督问答任务上取得显著的提升；Miao 等 (2016a) 使用变分自编码器实现了文本压缩的多样性；Zhang 等 (2016) 将变分自编码器引入神经机器翻译模型以帮助 RNNSearch 建立全局信息。

4.3　研究背景

本节简要介绍变分自编码器和条件变分自编码器的基本原理。

4.3.1　变分自编码器

变分自编码器（Variational Auto Encoder, VAE）首先被提出用于图像生成 (Kingma 等, 2014)，其主要目的是建立一个近似的后验分布 $q_\phi(z|y)$，其训练目标为最大化 $\log p_\theta(y)$：

$$\begin{aligned}\log p_\theta(y) &\geq \int \log p(y|z)p(z)dz \\ &= \mathrm{E}_{q_\phi(z|y)}[\log \frac{p(y|z)p(z)}{q(z|y)}] \\ &= -\mathrm{KL}(q_\phi(z|y)||p_\theta(z)) + \mathrm{E}_{q_\phi(z|y)} \log p_\theta(y|z) \\ &= \mathcal{L}(\theta, \phi; y)\end{aligned} \tag{4.1}$$

该目标函数也被称为 Evidence Lower Bound (ELBO)。

4.3.2　条件变分自编码器

条件变分自编码器（Conditional Variational Auto Encoder, CVAE）首先被提出用于特定条件下的图像生成 (Sohn 等, 2015)。条件变分自编码器是变分自编码器在给定条件下的拓展，它的目标建立给定条件 x 下标签 y 的隐变量 z 的后验分布（Posterior）$q_\phi(z|x, y)$，其训练目标为最大化 $\log p_\theta(y|x)$：

$$\begin{aligned}\log p_\theta(y|x) &\geq \int \log p(y|x, z)p(z|x)dz \\ &= -\mathrm{KL}(q_\phi(z|x, y)||p_\theta(z|x)) + \mathrm{E}_{q_\phi(z|x,y)} \log p_\theta(y|x, z) \\ &= \mathcal{L}(\theta, \phi; x, y)\end{aligned} \tag{4.2}$$

对于机器翻译来说，给定条件为 $(\mathbf{x}, \mathbf{y}_{<\mathbf{t}})$，标签为 y_t，本章希望 CVAE 为标签 y_t 建立一个近似的隐变量分布 $q_\phi(z|\mathbf{x}, \mathbf{y}_{<\mathbf{t}})$ 用于标签平滑。

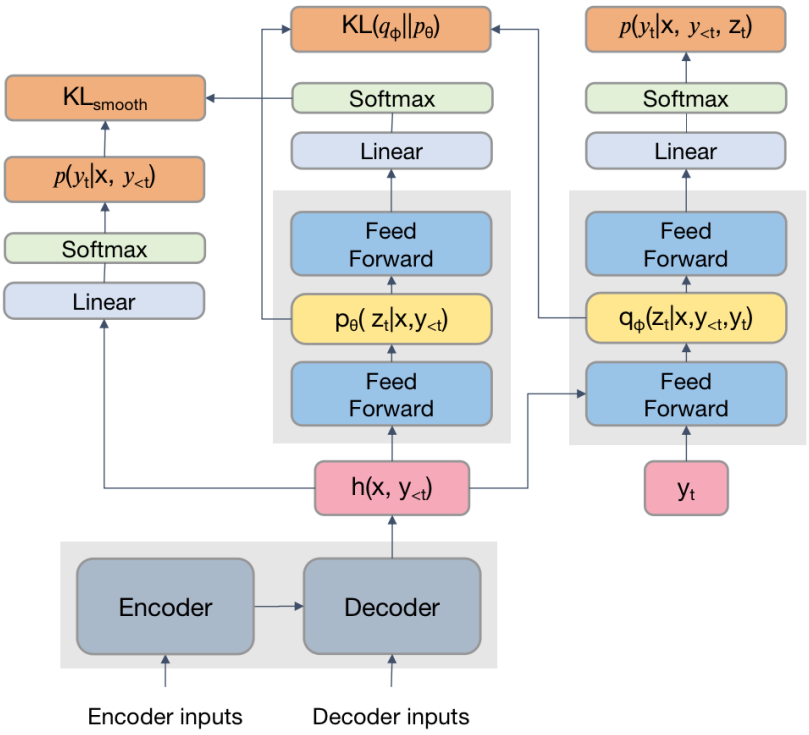


图 4.1 CVAE 标签平滑模型的结构示意图

Figure 4.1 The Structure of CVAE label smoothing model

4.4 基于条件变分自编码器的标签平滑

本章希望利用条件变分自编码器建模给定条件下词级别标签的隐变量分布，并从该分布采样出新的标签用于神经机器翻译模型的标签平滑。

如图4.1所示，本章在 Transformer 模型的基础上设计添加了一个 CVAE 模块，该 CVAE 模块分为先验部分和后验部分，分别建立了 CVAE 的先验分布和后验分布，本章设定两个分布均为高斯分布。其中，先验分布 p_θ 的方差和均值由 Transformer 解码器的输出 $h_{\mathbf{x},\mathbf{y}_{<t}}$ 经过得到 FFN 网络和线性映射得到，而后验分布 q_ϕ 的均值和方差由 $h_{\mathbf{x},\mathbf{y}_{<t}}$ 和 y_t 经过 FFN 网络和线性映射得到。本章利用重参数化技巧从后验分布重构出标签 y_t，并从先验分布生成标签用于神经机器翻译模型的标签平滑。

4.4.1 建立先验分布

条件变分自编码器（CVAE）的先验分布建立在“条件”的基础上。对于机器翻译任务来说，该条件是编码器的输入 X 和解码器的输入 $y_{<t}$。在实验中，本

章使用解码器输出的隐状态 $h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}}$ 作为条件建立条件变分自编码器模块的先验分布。

为了增强 CVAE 模块的学习能力，本章使用 Transformer 模型中的前馈神经网络对解码器输出的隐状态 $h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}}$ 进行处理。在该层前馈神经网络中，本章使用 tanh 函数作为激活函数，并使用了残差连接（Residual Connection) 和层规范 (Layer Normalization) 的技术。在前馈神经网络的基础上，本章通过两个不同的线性映射，分别得到先验分布的方差 σ_{prior} 和均值 μ_{prior}：

$$c_{\text{prior}} = \text{layernorm}(\tanh(h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}}\text{W}_1)\text{W}_2 + h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}}), \tag{4.3}$$

$$\log \sigma_{\text{prior}} = c_{\text{prior}}\text{W}_{\sigma 1}, \tag{4.4}$$

$$\mu_{\text{prior}} = c_{\text{prior}}W_{\mu 1}, \tag{4.5}$$

其中 c_{prior} 是前馈神经网络的输出，W_1 和 W_2 是前馈神经网络的线性映射矩阵，大小分别是 512×2048 和 2048×512，$\text{W}_{\sigma 1}$ 和 $\text{W}_{\mu 1}$ 是两个大小为 512×512 的线性映射矩阵。

4.4.2　建立后验分布

条件变分自编码器的后验分布是由“条件”和“标签”共同建立的，对于机器翻译任务来说，后验分布是基于模型的输入 X 、$y_{<t}$ 和标签 y_t 建立的。和先验分布的建立方式类似，本章使用前馈神经网络和线性映射建立后验分布的方差 $\sigma_{\text{posterior}}$ 和均值 $\mu_{\text{posterior}}$。

首先，本章通过线性映射将 X 、$y_{<t}$、y_t 转化为 512 维：

$$h_{\text{posterior}} = [h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}} : y_t]\text{W}_3, \tag{4.6}$$

其中 [:] 表示将两个 512 维的向量拼接成 1024 维，W_3 是 1024×512 的线性映射矩阵，$h_{\text{posterior}}$ 是 512 维的隐状态。在该基础上，本章建立后验分布：

$$c_{\text{posterior}} = \text{layernorm}(\tanh(h_{\mathbf{x},\mathbf{y}_{<\mathbf{t}}}\text{W}_3)\text{W}_4 + h_{\text{posterior}}), \tag{4.7}$$

$$\log \sigma_{\text{posterior}} = c_{\text{posterior}}\text{W}_{\sigma 2}, \tag{4.8}$$

$$\mu_{\text{posterior}} = c_{\text{posterior}}\text{W}_{\mu 2}, \tag{4.9}$$

其中 $c_{\text{posterior}}$ 是前馈神经网络的输出，W_3 和 W_4 是前馈神经网络的线性映射矩阵，大小分别是 512×2048 和 2048×512，$\text{W}_{\sigma 2}$ 和 $\text{W}_{\mu 2}$ 是两个大小为 512×512 的线性映射矩阵。

4.4.3 条件变分自编码器的解码器

条件变分自编码器的隐变量需要通过解码器才能转化为对目标端词的预测分布。对于从先验分布采样的随机变量 z_t^{θ}，本章通过前馈神经网络和对词表的线性映射建立预测分布：

$$s_t = \text{layernorm}(\tanh(z_t^{\theta}\text{W}_5)\text{W}_6 + z_t^{\theta}), \tag{4.10}$$

$$p((y_t|z_t^{\theta}) = \text{softmax}(\text{W}_\text{V} s_t^{\text{T}}), \tag{4.11}$$

其中 W_5 和 W_6 是前馈神经网络的线性映射矩阵，大小分别是 512×2048 和 2048×512，s_t 是条件变分自编码器输出的隐状态，W_V 代表神经机器翻译模型的输出层词向量参数矩阵，大小为 $|\text{V}| \times 512$，V 是目标端词表。

相似地，对于从后验分布采样的随机变量 z_t^{ϕ}，其重构的预测分布为：

$$s_t = \text{layernorm}(\tanh(z_t^{\phi}\text{W}_5)\text{W}_6 + z_t^{\phi}), \tag{4.12}$$

$$p((y_t|\text{X}, y_{<t}, z_t^{\phi}) = \text{softmax}(\text{W}_\text{V}(s_t + h_{\mathbf{x},\mathbf{y_{<t}}})^{\text{T}}), \tag{4.13}$$

其中 $h_{\mathbf{x},\mathbf{y_{<t}}}$ 是 Transformer 模型解码器的输出，作为自编码的条件。

4.4.4 损失函数

图4.1中顶部的橙色模块为模型的四个损失函数项。首先，条件变分自编码器的损失函数由重构损失和先验后验之间的 KL 距离组成（ELBO）：

$$\begin{aligned}\mathcal{L}_{\text{CVAE}}(\text{X}, \text{Y}) = -\frac{1}{\text{J}}\sum_{t=1}^{\text{J}+1}[&\log p(y_t|\text{X}, y_{<t}, z_t^{\phi}) \\ &-\text{KL}(q_{\phi}(z_t|\text{X}, y_{<t}, y_t)||p_{\theta}(z_t|\text{X}, y_{<t})].\end{aligned} \tag{4.14}$$

其中 J 为目标句的句长，z_t^{ϕ} 表示标签 $p(y_t|\text{X}, y_{<t}, z_t^{\phi})$ 是由后验分布 p_{ϕ} 重参数化采样得到的。

其次，CVAE 通过先验分布重采样得到的预测分布，被实时地用于机器翻译模型的标签平滑：

$$\begin{aligned}\mathcal{L}_{\text{MT}}(\text{X}, \text{Y}) = -\frac{1}{\text{J}}\sum_{t=1}^{\text{J}+1}[&(1-\alpha)\log p(y_t|\text{X}, y_{<t}) \\ &-\alpha\text{KL}(p(y_t|\text{X}, y_{<t})||p(y_t|\text{X}, y_{<t}, z_t^{\theta})],\end{aligned} \tag{4.15}$$

其中 α 是控制标签平滑比重的超参，z_t^θ 表示标签 $p(y_t|\mathrm{X}, y_{<t}, z_t^\theta)$ 是由先验分布 p_θ 重参数化采样得到的。由于该 KL 项并不向 CVAE 传递梯度，因此相当于标签平滑。

最后，模型的总损失函数为机器翻译和 CVAE 模型损失函数的加权和：

$$\mathcal{L}(\mathrm{X}, \mathrm{Y}) = (1-\beta)\mathcal{L}_{\mathrm{MT}}(\mathrm{X}, \mathrm{Y}) + \beta\mathcal{L}_{\mathrm{CVAE}}(\mathrm{X}, \mathrm{Y}), \tag{4.16}$$

其中 β 是控制 NMT 和 CVAE 模型损失函数比重的超参。

4.4.5　重参数化技巧

在训练条件变分自编码器时，我们需要从后验分布中采样并重构出标签；在生成平滑标签时，我们需要从先验分布采样并生成平滑标签。直接从隐变量分布采样会导致梯度无法回传，为此研究者们提出了重参数化技巧（Reparameterization trick）。具体而言，我们可以利用高斯分布的性质，先从标准的高斯分布采样出随机变量，再使用均值和方差映射到对应的分布：

$$x_1 \sim \mathrm{N}(0, 1), \tag{4.17}$$

$$x_2 \sim \mathrm{N}(0, 1), \tag{4.18}$$

$$z_{prior} = \sigma_{prior} x_1 + \mu_{prior}, \tag{4.19}$$

$$z_{posterior} = \sigma_{posterior} x_2 + \mu_{posterior}, \tag{4.20}$$

其中 $z_{prior} \sim \mathrm{N}(\mu_{prior}, \sigma_{prior})$ 和 $z_{posterior} \sim \mathrm{N}(\mu_{posterior}, \sigma_{posterior})$ 分别是条件变分自编码器先验分布和后验分布的随机变量。这样，随机隐变量 z_{prior} 和 $z_{posterior}$ 的梯度可以通过分布的方差和均值继续回传。

4.4.6　缓解后验坍塌的门控机制

条件变分自编码的重构误差 $\log p(y_t|\mathrm{X}, y_{<t}, z_t^\phi)$ 中包含自编码条件 $(\mathrm{X}, y_{<t})$。由于神经机器翻译模型的拟合能力很强，重构误差可能会过于依赖条件 $(\mathrm{X}, y_{<t})$，而忽略了对后验分布的学习，形成后验坍塌（Posterior Collapse）。为了缓解这个问题，本章对公式 4.13 添加了一个门控机制：

$$p((y_t|\mathrm{X}, y_{<t}, z_t^\phi) = \mathrm{softmax}(\mathrm{W_V}(\lambda s_t + (1-\lambda)h_{\mathbf{x},\mathbf{y_{<t}}})^{\mathrm{T}}), \tag{4.21}$$

其中 λ 是控制条件信息和后验信息比重的超参数。

4.5 实验

本章在两个常用的数据集对 CVAE 标签平滑进行了实验，并把它和相关的对比方法进行了比较。

4.5.1 数据集

本章分别在 WMT16 英罗和 NIST 中英进行了实验。

- WMT16 英文-罗马尼亚文数据集包含大约 60 万行平行语料，由 Moses 工具 (Koehn 等, 2007b) 进行 tokenize，词表进行了 4 万次合并操作的 BPE (Sennrich 等, 2016b) 处理。本章分别使用 news-dev 2016 和 news-test 2016 作为验证集和测试集。
- NIST 中文-英文数据集包含 125 万行平行语料，其中英文数据使用了 Moses 进行 tokenize，中文数据使用了斯坦福分词工具进行 tokenize。数据集的词表经过了 3.2 万次合并操作的 BPE 处理。本章使用 mt02、mt03、mt04、mt05、mt06、mt08 数据集作为测试集。

4.5.2 参数设置

为了公平地比较所有方法，本章在 Transformer (Vaswani 等, 2017) 模型上对所有方法进行了复现。本章的实验都是在开源工具 Fairseq (Ott 等, 2019) 的基础上进行的，使用了完全相同的实验配置和硬件设备。

本章的实验使用了 Transformer base 的实验配置，编码器和解码器都是 6 层，模型维度为 512 维，注意力机制使用了 8 个头（Head），前馈网络为 2048 维。所有模型都使用了 4000 个预热步（warm-up steps），初始学习率为 $7e^{-4}$，标签平滑比重为 0.1，Dropout 概率为 0.1，Adam 的参数为 $\beta_1 = 0.9$、$\beta_2 = 0.98$ 和 $\epsilon = 1e^{-9}$(Vaswani 等, 2017)。所有的实验都是在 4 块 GeForce RTX 3090 图像处理器进行的，批量训练的大小（Batch size）为 4096×8 个 token。在测试时，柱搜索参数（Beam Size）为 4，长度惩罚（Length Penalty）为 0.6。

对于英罗数据集，本章共享了源端和目标端词表，词表大小均为 32768，在训练时使用了 TWWT (Press 等, 2017) 的设置，模型一共训练了 50 轮。对于中英数据集，中文和英文的词表大小分别为 40072 和 29408，模型一共训练了 35 轮。对于超参数 α 、β 和 λ，英罗数据集的设置为 $(\alpha = 0.3, \beta = 0.35, \lambda = 1)$，中英数据集的设置为 $(\alpha = 0.35, \beta = 0.4, \lambda = 1)$。

4.5.3　对比方法

本章复现了四种在线的标签平滑方法和一种离线的标签平滑方法。“在线”是指标签生成器是与机器翻译模型同时训练的，而“离线”是指标签生成器是单独预训练。

- 虚假标签平滑（Fake Label）。虚假标签平滑 (Luo 等, 2021) 认为使用平均分布进行标签平滑的做法虽然有效，但是其做法并不合理，因为词表中大多数词与标签是不相关的。为此，作者提出一个替换性的方法，即降低硬标签的概率，降低的概率值并不分给词表里的其他词，而是只分给了一个虚假标签（Fake Label）。该标签只用于模型的标签平滑，不出现在数据集中。
- 类别相似度标签平滑（Class-Similarity Label, Simi Label）(Liu 等, 2021)。类别相似度标签将分类任务中类别的相似性作为平滑标签，具体在机器翻译任务上，本章使用类别特征向量的相似度来衡量类别相似度。
- 历史标签平滑（Epoch Label）(Zhang 等, 2021)。该方法提出将每轮训练中模型产生的属于同一标签的预测分布的平均，可以作为该标签在下轮训练的平滑标签，因此该方法可以看作一种在线标签平滑方法。实验上，由于机器翻译的标签数量比较庞大（3-4 万），训练时需要维持并更新一个巨大的张量，因此训练成本是基线模型的 8-9 倍。
- 自蒸馏。本章复现了离线知识蒸馏（Offline-Distill）(Furlanello 等, 2018) 和在线知识蒸馏（Onlline-Distill）(Yuan 等, 2020)。离线自蒸馏是指先训练好一个模型，然后再使用该模型蒸馏另一个从新训练的模型。在线自蒸馏是指使用模型在当前训练步预测的标签作为平滑标签。

4.5.4　实验结果

各标签平滑方法在英罗数据集和中英数据集上的效果如图4.1和图4.2所示。首先本章注意到，在线自蒸馏在机器翻译上几乎没有效果。另外，虚假标签平滑（Fake-Label）、类别相似度标签平滑（Simi-Label）和历史标签平滑（Epoch-Label）的提升并不明显，最高在 0.5 BLEU 左右。尤其是历史标签平滑，在训练时间、空间成本都很高的情况下提升却很小。以上模型都属于在线标签平滑的范畴，它们的提升都比较小，这体现了在线标签平滑的难度。

作为一种在线的标签平滑方法，CVAE 在英罗数据集和中英数据集上都取得

表 4.1 在 WMT16 英文-罗马尼亚文数据集上的实验结果

Table 4.1 Experimental results on WMT16 En-Ro task

	WMT2016 En→Ro		
	2016	Δ	time
Transformer	32.62	–	1.00x
Offline-Distill	**34.72**	**+2.00**	**3.27x**
Onlline-Distill	33.69	+ 0.07	1.00x
Fake-Label	32.73	+ 0.11	1.00x
Simi-Label	33.24	+ 0.62	1.07x
Epoch-Label	32.98	+ 0.36	8.78x
CVAE-Label	**32.83**	**+1.21**	**1.23x**

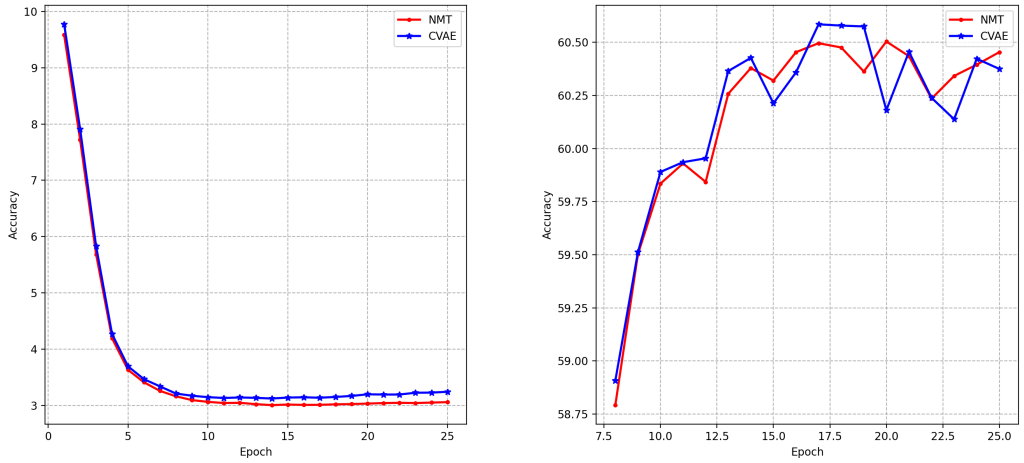
表 4.2 在 NIST 中文-英文数据集上的实验结果

Table 4.2 Experimental results on NIST Zh-En task

	NIST Zh-En							
	mt02	mt03	mt04	mt05	mt06	mt08	Avg	time
Transformer	45.31	44.13	45.82	43.9	43.72	34.9	42.96	1.00x
Offline-Distill	47.22	46.06	47.83	45.79	45.85	36.77	**44.90**	**3.21x**
Online-Distill	45.54	44.01	45.96	44.21	43.65	34.95	43.05	1.00x
Fake-Label	45.46	44.21	45.99	44.55	43.80	34.33	43.06	1.00x
Simi-Label	45.45	43.89	46.06	44.55	44.41	35.28	43.27	1.07x
Epoch-Label	45.61	44.65	45.91	44.09	44.58	34.98	43.30	8.76x
CVAE-Label	45.67	45.82	46.65	46.30	45.16	35.36	**44.16**	**1.22x**

了 1.2 BLEU 左右的提升，远远超过了已有的在线标签平滑方法。离线自蒸馏虽然取得了 2 个 BLEU 左右的提升，但是却需要 3 倍以上的训练时间和 2 倍的参数量。CVAE 标签平滑方法作为一种轻量的在线的标签平滑方法，具有较小的训练成本和显著的性能提升，在成本和性能之间取得了更好的平衡。

4.5.5　CVAE 模块的训练情况



(a)　　**(b)**

图 4.2 NMT 和 CVAE 模块在测试集上的预测损失和预测准确率

Figure 4.2 Prediction loss and prediction accuracy of NMT and CVAE modules on the test set.

CVAE 模块生成的标签应该同时具备准确性和多样性，为了验证这一点，本节在 NIST 中英数据集上分析了 CVAE 模块的训练情况。如图 4.2 (a) 所示，在训练过程中，CVAE 模块在测试集上的重构损失始终高于神经机器翻译模型；图 4.2 (b) 表明，CVAE 模块在测试集的准确率和神经机器翻译模型不相上下。以上两点说明 CVAE 模块的对正确标签的预测准确率不亚于神经机器翻译模型，但是对硬标签（Hard Label）过拟合程度要低于神经机器翻译模型，因此 CVAE 模块重采样的标签具有较高的准确率和较低的过拟合程度，能够满足标签平滑的要求，这符合本章的设计目标。

4.5.6　先验分布和后验分布的消融实验

CVAE 模块学习了两个隐变量的分布，分别是先验分布和后验分布。按照 CVAE 的惯例，平滑标签应当使用先验分布来生成，但后验分布生成的标签同样

表 4.3 关于先验分布和后验分布的消融实验

Table 4.3 Ablation study about prior and post distribution

	WMT2016 En→Ro	
	2016	Δ
Transformer	32.62	–
先验	**32.83**	**+1.21**
后验	33.81	+ 1.19
先验 + 后验	32.81	+ 1.19

具备标签平滑的能力。为此，本节对两个分布进行了消融实验。如表 4.3 所示，使用 CVAE 的先验分布和后验分布生成平滑标签的差异很小，本章认为这主要有以下两个原因：（1）先验分布是通过神经网络建立的，而不是人为设定的固定值，因此更容易拟合。（2）神经机器翻译模型的参数量很大，拟合能力很强，因此能够将 CVAE 损失函数中的 KL 距离降低得很小，导致学习的先验分布和后验分布之间的差异很小。

4.6 本章小结

由于训练数据的相对稀缺和使用硬标签（Hard Label）作为拟合目标的训练方式，神经机器翻译模型容易过拟合训练集中出现的标签，而忽略其他可能正确的标签。为此，Szegedy 等 (2016) 提出标签平滑方法，使用平均分布和硬标签的加权训练模型，但其将硬标签的概率平均分配给所有其它标签的做法不符合真实的语言场景。本章提出一种轻量、高效的在线标签平滑方法以缓解模型对训练集标签过度自信的问题。本章利用条件变分自编码器能够基于给定条件将数据标签编码到一个隐变量空间并重构出来的特点，为神经机器翻译设计了一个基于条件变分自编码器的平滑标签生成器。该生成器能够学习给定源端和目标端输入条件下的词级别标签的隐变量分布，并通过该分布实时采样隐变量并生成新标签，用于神经机器翻译模型的在线标签平滑。实验上，条件变分自编码器模块和神经机器翻译一起训练，具有轻量、高效、在线的特点，能够较好地缓解模型对训练标签过度自信的问题。

在 NIST 中文-英文数据集和 WMT 英文-罗马尼亚文数据集上，本章的方法

取得了 1.20 BLEU 左右的提升，显著超过了已有的对比方法。分析实验表明，CVAE 模块能够生成具有较高准确率和较低自信度的标签，较好地满足了标签平滑的需求。

第 5 章　总结与展望

5.1　总结

神经机器翻译模型具有很强的性能，但也存在容易过拟合的缺点，模型往往在训练集上表现得很好，却在测试集上表现很差；另外，模型对噪声十分敏感，在输入中添加微小的噪声会导致模型的性能出现大幅度的下降。为了缓解神经机器翻译模型的过拟合问题，研究者们提出了一系列正则化方法，包括约束模型词向量参数的词向量正则化方法、提高模型噪声鲁棒性的扰动正则化方法和优化模型训练标签的标签平滑方法等。这些方法取得了显著的效果，但也分别存在一些需要改进的问题。

为了进一步解决神经机器翻译模型的过拟合问题，本文针对词向量正则化、扰动正则化和标签平滑三个研究内容，分别提出了以下三个方法：

（1）为了解决神经机器翻译模型词向量参数冗余度过高、语义特征较差的问题，本文提出一种增强词向量语义相关性的正则化方法，该方法完全基于双语语料的内在特征，适用于所有翻译方向且不依赖外部的知识指导。具体而言，机器翻译的一个训练样本由一个源端句子和一个目标端句子组成，其中隐含了单语言的词共现信息和双语言之间的词共现信息。基于这两种共现信息，本文提出一个自编码的训练目标来同时促进词向量的单语相关性和对齐相关性。实现上，本文借鉴 Skip-gram 的思想，将平行句对看作一个整体的序列，让其中的每个词预测序列中的其他词，从而使得频繁共现在平行句对的词学习到相近的词向量。实验结果表明，语义相关正则化能够极大地增强词向量参数的语义特征并显著提升神经机器翻译模型的性能。

（2）为了更好地提升模型对噪声的鲁棒性，本文对扰动正则化方法的工作原理进行了分析和改进。扰动正则化方法的前提假设是模型对原始样本的拟合程度总是高于对被扰动样本的拟合程度。本文利用模型在输入受到扰动前后对目标词的预测差异分析了词级别样本的拟合情况，发现该前提并不总是成立，模型对相当一部分训练样本是相对欠拟合的，对这部分样本进行扰动正则化训练会损害模型的性能。为了同时缓解模型的过拟合和欠拟合问题，本文提出预测差异正则化（PD-R），将输入扰动引起的预测差异作为正则项训练神经机器翻译模

型。在常用的数据集上，本文的方法相比现有方法取得了显著的提升，进一步提升了模型的泛化性能和对噪声的鲁棒性。

（3）为了解决神经机器翻译模型对训练标签过度自信的问题，本文提出了一种轻量、高效的在线标签平滑方法。本文利用条件变分自编码器能够基于给定条件将数据标签编码到一个隐变量空间并重构出来的特点，为神经机器翻译设计了一个基于条件变分自编码器的平滑标签生成器。该生成器能够学习给定源端和目标端输入条件下的词级别标签的隐变量分布，并通过该分布实时采样隐变量并生成新标签，用于神经机器翻译模型的在线标签平滑。实验上，条件变分自编码器能够和神经机器翻译一起训练，具有轻量、高效、在线的特点。实验结果表明，本文的方法能够较好地缓解模型对训练标签过度自信的问题并提升模型的翻译性能。

本文的三个正则化方法围绕过拟合问题，针对模型复杂度过高、训练数据稀疏和训练目标不合理这三个引起过拟合的原因分别提出了对应的解决方案。其中词向量正则化方法从参数层面入手，利用平行语料的词共现信息提升了词向量参数的语义特征，约束了词向量参数的表达自由度；扰动正则化方法从数据层面入手，通过加噪训练干扰了模型对训练数据的拟合，提升了模型在噪声数据上的表现；标签平滑方法从标签层面入手，对模型的训练标签进行了平滑，缓解了模型对训练标签的过度自信问题。在原理上，三个方法相互独立，解决的问题和提出的方法没有重叠；在实验上，三个方法能够互补，在应用时可以叠加使用以最大化地提高模型的性能。

5.2 展望

神经机器翻译和深度学习永远面临着数据稀缺和过拟合的问题，正则化方法的研究也永远不会停息，当前方法以及本文提出的三个方法必然会在不远的未来被新方法所超越。我很期待在未来随着人们对神经网络模型的认识不断深入，研究者们能够推陈出新，提出更加简洁有效的正则化方法。

参考文献

Ataman D, Federico M. Compositional representation of morphologically-rich input for neural machine translation [C/OL]//Gurevych I, Miyao Y. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers. Association for Computational Linguistics, 2018: 305-311. https://aclanthology.org/P18-2049/.

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C/OL]//Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. http://arxiv.org/abs/1409.0473.

Barone A V M, Haddow B, Germann U, et al. Regularization techniques for fine-tuning in neural machine translation [C/OL]//Palmer M, Hwa R, Riedel S. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. Association for Computational Linguistics, 2017: 1489-1494. https://doi.org/10.18653/v1/d17-1156.

Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation [C/OL]// 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=BJ8vJebC-.

Bengio S, Vinyals O, Jaitly N, et al. Scheduled sampling for sequence prediction with recurrent neural networks [C/OL]//Cortes C, Lawrence N D, Lee D D, et al. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015: 1171-1179. https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Berend G. Sparse coding of neural word embeddings for multilingual sequence labeling [J/OL]. Trans. Assoc. Comput. Linguistics, 2017, 5: 247-261. https://transacl.org/ojs/index.php/tacl/article/view/1063.

Berend G. ℓ_1 regularization of word embeddings for multi-word expression identification [J/OL]. Acta Cybern., 2018, 23(3): 801-813. https://doi.org/10.14232/actacyb.23.3.2018.5.

Bowman S R, Vilnis L, Vinyals O, et al. Generating sentences from a continuous space [J/OL]. CoRR, 2015, abs/1511.06349. http://arxiv.org/abs/1511.06349.

Briakou E, Carpuat M. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation [C/OL]//Proceedings of the 59th Annual Meeting of the Association

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 7236-7249. https://aclanthology.org/2021.acl-long.562.

Brown P F, Cocke J, Pietra S D, et al. A statistical approach to machine translation [J]. Comput. Linguistics, 1990, 16(2): 79-85.

Brown P F, Pietra S D, Pietra V J D, et al. The mathematics of statistical machine translation: Parameter estimation [J]. Comput. Linguistics, 1993, 19(2): 263-311.

Caglayan O, Barrault L, Bougares F. Multimodal attention for neural machine translation [J/OL]. CoRR, 2016, abs/1609.03976. http://arxiv.org/abs/1609.03976.

Calixto I, Liu Q. Incorporating global visual features into attention-based neural machine translation [C/OL]//Palmer M, Hwa R, Riedel S. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. Association for Computational Linguistics, 2017: 992-1003. https://doi.org/10.18653/v1/d17-1105.

Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C/OL]// Vedaldi A, Bischof H, Brox T, et al. Lecture Notes in Computer Science: volume 12346 Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Springer, 2020: 213-229. https://doi.org/10.1007/978-3-030-58452-8_13.

Chen W, Grangier D, Auli M. Strategies for training large vocabulary neural language models [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016. https://doi.org/10.18653/v1/p16-1186.

Cheng Y, Shen S, He Z, et al. Agreement-based joint training for bidirectional attention-based neural machine translation [C/OL]//Kambhampati S. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. IJCAI/AAAI Press, 2016: 2761-2767. http://www.ijcai.org/Abstract/16/392.

Cheng Y, Yang Q, Liu Y, et al. Joint training for pivot-based neural machine translation [C/OL]// Sierra C. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. ijcai.org, 2017: 3974-3980. https://doi.org/10.24963/ijcai.2017/555.

Cho K, Esipova M. Can neural machine translation do simultaneous translation? [J/OL]. CoRR, 2016, abs/1606.02012. http://arxiv.org/abs/1606.02012.

Cho K, van Merrienboer B, Gülçehre Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C/OL]//Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP

2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2014: 1724-1734. https://doi.org/10.3115/v1/d14-1179.

Chu C, Dabre R, Kurohashi S. An empirical comparison of domain adaptation methods for neural machine translation [C/OL]//Barzilay R, Kan M. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers. Association for Computational Linguistics, 2017: 385-391. https://doi.org/10.18653/v1/P17-2061.

Costa-jussà M R, Fonollosa J A R. Character-based neural machine translation [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers. The Association for Computer Linguistics, 2016. https://doi.org/10.18653/v1/p16-2058.

Dalvi F, Durrani N, Sajjad H, et al. Incremental decoding and training methods for simultaneous translation in neural machine translation [C/OL]//Walker M A, Ji H, Stent A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, 2018: 493-499. https://doi.org/10.18653/v1/n18-2079.

Demeter D, Kimmel G, Downey D. Stolen probability: A structural weakness of neural language models [C/OL]//Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020: 2191-2197. https://doi.org/10.18653/v1/2020.acl-main.198.

Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C/OL]//Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 4171-4186. https://doi.org/10.18653/v1/n19-1423.

Ding Y, Liu Y, Luan H, et al. Visualizing and understanding neural machine translation [C/OL]// Barzilay R, Kan M. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. Association for Computational Linguistics, 2017: 1150-1159. https://doi.org/10.18653/v1/P17-1106.

Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation [C/OL]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation

of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. The Association for Computer Linguistics, 2015: 1723-1732. https://doi.org/10.3115/v1/p15-1166.

Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [C/OL]//9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. https://openreview.net/forum?id=YicbFdNTTy.

Dubey A, Gupta O, Guo P, et al. Pairwise confusion for fine-grained visual classification [C/OL]// Ferrari V, Hebert M, Sminchisescu C, et al. Lecture Notes in Computer Science: volume 11216 Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII. Springer, 2018: 71-88. https://doi.org/10.1007/978-3-030-01258-8_5.

Duckworth D, Neelakantan A, Goodrich B, et al. Parallel scheduled sampling [J/OL]. CoRR, 2019, abs/1906.04331. http://arxiv.org/abs/1906.04331.

Ebrahimi J, Rao A, Lowd D, et al. HotFlip: White-box adversarial examples for text classification [C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 31-36. https://aclanthology.org/P18-2006.

Fang Q, Ye R, Li L, et al. STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022.

Feng Y, Zhang S, Zhang A, et al. Memory-augmented neural machine translation [C/OL]//Palmer M, Hwa R, Riedel S. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. Association for Computational Linguistics, 2017: 1390-1399. https://doi.org/10.18653/v1/d17-1146.

Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism [C/OL]//Knight K, Nenkova A, Rambow O. NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. The Association for Computational Linguistics, 2016: 866-875. https://doi.org/10.18653/v1/n16-1101.

Freitag M, Al-Onaizan Y. Fast domain adaptation for neural machine translation [J/OL]. CoRR, 2016, abs/1612.06897. http://arxiv.org/abs/1612.06897.

Furlanello T, Lipton Z C, Tschannen M, et al. Born-again neural networks [C/OL]//Dy J G, Krause A. Proceedings of Machine Learning Research: volume 80 Proceedings of the 35th International

Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. PMLR, 2018: 1602-1611. http://proceedings.mlr.press/v80/furlanello18a.html.

Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks [C/OL]//Lee D D, Sugiyama M, von Luxburg U, et al. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. 2016: 1019-1027. https://proceedings.neurips.cc/paper/2016/hash/076a0c97d09cf1a0ec3e19c7f2529f2b-Abstract.html.

Garg S, Peitz S, Nallasamy U, et al. Jointly learning to align and translate with transformer models [C/OL]//Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, 2019: 4452-4461. https://doi.org/10.18653/v1/D19-1453.

Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [C/OL]// Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. http://arxiv.org/abs/1412.6572.

Gu J, Neubig G, Cho K, et al. Learning to translate in real-time with neural machine translation [C/OL]//Lapata M, Blunsom P, Koller A. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers. Association for Computational Linguistics, 2017: 1053-1062. https://doi.org/10.18653/v1/e17-1099.

Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=B1l8BtlCb.

Gu J, Tresp V. Contextual prediction difference analysis [J/OL]. CoRR, 2019, abs/1910.09086. http://arxiv.org/abs/1910.09086.

Guo D, Kim Y, Rush A. Sequence-level mixed sample data augmentation [C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020a: 5547-5552. https://aclanthology.org/2020.emnlp-main.447.

Guo F, Zhao Q, Li X, et al. Detecting adversarial examples via prediction difference for deep neural networks [J/OL]. Inf. Sci., 2019, 501: 182-192. https://doi.org/10.1016/j.ins.2019.05.084.

Guo J, Tan X, Xu L, et al. Fine-tuning by curriculum learning for non-autoregressive neural machine translation [C/OL]//The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020,

The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 2020b: 7839-7846. https://ojs.aaai.org/index.php/AAAI/article/view/6289.

Ha T, Niehues J, Waibel A. Toward multilingual neural machine translation with universal encoder and decoder [C/OL]//Proceedings of the 13th International Conference on Spoken Language Translation, IWSLT 2016, Seattle, WA, USA, December 8-9, 2016. International Workshop on Spoken Language Translation, 2016. https://aclanthology.org/2016.iwslt-1.6.

Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J/OL]. CoRR, 2012, abs/1207.0580. http://arxiv.org/abs/1207.0580.

Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network [J/OL]. CoRR, 2015, abs/1503.02531. http://arxiv.org/abs/1503.02531.

Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 6525-6538. https://doi.org/10.18653/v1/p19-1653.

Jean S, Lauly S, Firat O, et al. Does neural machine translation benefit from larger context? [J/OL]. CoRR, 2017, abs/1704.05135. http://arxiv.org/abs/1704.05135.

Johnson M, Schuster M, Le Q V, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation [J/OL]. Trans. Assoc. Comput. Linguistics, 2017, 5: 339-351. https://transacl.org/ojs/index.php/tacl/article/view/1081.

Kalchbrenner N, Blunsom P. Recurrent continuous translation models [C/OL]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2013: 1700-1709. https://aclanthology.org/D13-1176/.

Karpukhin V, Levy O, Eisenstein J, et al. Training on synthetic noise improves robustness to natural noise in machine translation [C/OL]//Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong, China: Association for Computational Linguistics, 2019: 42-47. https://aclanthology.org/D19-5506.

Khayrallah H, Koehn P. On the impact of various types of noise on neural machine translation [C/OL]//Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. Melbourne, Australia: Association for Computational Linguistics, 2018: 74-83. https://aclanthology.org/W18-2709.

Kim Y, Rush A M. Sequence-level knowledge distillation [C/OL]//Su J, Carreras X, Duh K. Pro-

ceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. The Association for Computational Linguistics, 2016: 1317-1327. https://doi.org/10.18653/v1/d16-1139.

Kingma D P, Welling M. Auto-encoding variational bayes [C/OL]//Bengio Y, LeCun Y. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014. http://arxiv.org/abs/1312.6114.

Kiros R, Zhu Y, Salakhutdinov R, et al. Skip-thought vectors [C/OL]//Cortes C, Lawrence N D, Lee D D, et al. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015: 3294-3302. https://proceedings.neurips.cc/paper/2015/hash/f442d33fa06832082290ad8544a8da27-Abstract.html.

Koehn P, Och F J, Marcu D. Statistical phrase-based translation [C/OL]//Hearst M A, Ostendorf M. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003. The Association for Computational Linguistics, 2003. https://aclanthology.org/N03-1017/.

Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C/OL]//Carroll J A, van den Bosch A, Zaenen A. ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. The Association for Computational Linguistics, 2007a. https://aclanthology.org/P07-2045/.

Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C/OL]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Prague, Czech Republic: Association for Computational Linguistics, 2007b: 177-180. https://aclanthology.org/P07-2045.

Krogh A, Hertz J A. A simple weight decay can improve generalization [C/OL]//Moody J E, Hanson S J, Lippmann R. Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]. Morgan Kaufmann, 1991: 950-957. http://papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization.

Kuang S, Li J, Branco A, et al. Attention focusing for neural machine translation by bridging source and target embeddings [C/OL]//Gurevych I, Miyao Y. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics, 2018: 1767-1776. https://aclanthology.org/P18-1164/.

Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates [C/OL]//Gurevych I, Miyao Y. Proceedings of the 56th Annual Meeting

of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics, 2018a: 66-75. https://aclanthology.org/P18-1007/.

Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018b: 66-75.

Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=rkYTTf-AZ.

Le Q V, Mikolov T. Distributed representations of sentences and documents [C/OL]//JMLR Workshop and Conference Proceedings: volume 32 Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. JMLR.org, 2014: 1188-1196. http://proceedings.mlr.press/v32/le14.html.

Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation [J/OL]. Trans. Assoc. Comput. Linguistics, 2017, 5: 365-378. https://transacl.org/ojs/index.php/tacl/article/view/1051.

Li C, Liu C, Duan L, et al. Reconstruction regularized deep metric learning for multi-label image classification [J/OL]. IEEE Trans. Neural Networks Learn. Syst., 2020, 31(7): 2294-2303. https://doi.org/10.1109/TNNLS.2019.2924023.

Li J, Tu Z, Yang B, et al. Multi-head attention with disagreement regularization [C/OL]//Riloff E, Chiang D, Hockenmaier J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 2018: 2897-2903. https://doi.org/10.18653/v1/d18-1317.

Li J, Gao P, Wu X, et al. Mixup decoding for diverse machine translation [C/OL]//Moens M, Huang X, Specia L, et al. Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021. Association for Computational Linguistics, 2021: 312-320. https://doi.org/10.18653/v1/2021.findings-emnlp.29.

Li X, Li G, Liu L, et al. On the word alignment from neural machine translation [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019a: 1293-1303. https://doi.org/10.18653/v1/p19-1124.

Li Z, Lin Z, He D, et al. Hint-based training for non-autoregressive machine translation [C/OL]//

Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, 2019b: 5707-5712. https://doi.org/10.18653/v1/D19-1573.

Liang B, Li H, Su M, et al. Deep text classification can be fooled [C/OL]//Lang J. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, 2018: 4208-4215. https://doi.org/10.24963/ijcai.2018/585.

Liang X, Wu L, Li J, et al. R-drop: Regularized dropout for neural networks [J/OL]. CoRR, 2021, abs/2106.14448. https://arxiv.org/abs/2106.14448.

Ling W, Trancoso I, Dyer C, et al. Character-based neural machine translation [J/OL]. CoRR, 2015, abs/1511.04586. http://arxiv.org/abs/1511.04586.

Liu C, JáJá J F. Class-similarity based label smoothing for confidence calibration [C/OL]//Farkas I, Masulli P, Otte S, et al. Lecture Notes in Computer Science: volume 12894 Artificial Neural Networks and Machine Learning - ICANN 2021 - 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part IV. Springer, 2021: 190-201. https://doi.org/10.1007/978-3-030-86380-7_16.

Liu L, Utiyama M, Finch A M, et al. Neural machine translation with supervised attention [C/OL]// Calzolari N, Matsumoto Y, Prasad R. COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. ACL, 2016: 3093-3102. https://aclanthology.org/C16-1291/.

Liu X, Wong D F, Liu Y, et al. Shared-private bilingual word embeddings for neural machine translation [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 3613-3622. https://doi.org/10.18653/v1/p19-1352.

Luo Z, Xi Y, Mao X. Smoothing with fake label [C/OL]//Demartini G, Zuccon G, Culpepper J S, et al. CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021. ACM, 2021: 3303-3307. https://doi.org/10.1145/3459637.3482184.

Luong M, Le Q V, Sutskever I, et al. Multi-task sequence to sequence learning [C/OL]//Bengio Y, LeCun Y. 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. 2016. http://arxiv.org/abs/1511.06114.

Ma M, Huang L, Xiong H, et al. STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework [C/OL]//Korhonen A, Traum D R, Màrquez L.

Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 3025-3036. https://doi.org/10.18653/v1/p19-1289.

Madhyastha P S, Wang J, Specia L. Sheffield multimt: Using object posterior predictions for multimodal machine translation [C/OL]//Bojar O, Buck C, Chatterjee R, et al. Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017. Association for Computational Linguistics, 2017: 470-476. https://doi.org/10.18653/v1/w17-4752.

Meng Y, Huang J, Wang G, et al. Spherical text embedding [C/OL]//Wallach H M, Larochelle H, Beygelzimer A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 8206-8215. https://proceedings.neurips.cc/paper/2019/hash/043ab21fc5a1607b381ac3896176dac6-Abstract.html.

Mi H, Wang Z, Ittycheriah A. Supervised attentions for neural machine translation [C/OL]//Su J, Carreras X, Duh K. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. The Association for Computational Linguistics, 2016: 2283-2288. https://doi.org/10.18653/v1/d16-1249.

Miao Y, Blunsom P. Language as a latent variable: Discrete generative models for sentence compression [C/OL]//Su J, Carreras X, Duh K. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. The Association for Computational Linguistics, 2016a: 319-328. https://doi.org/10.18653/v1/d16-1031.

Miao Y, Yu L, Blunsom P. Neural variational inference for text processing [C/OL]//Balcan M, Weinberger K Q. JMLR Workshop and Conference Proceedings: volume 48 Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR.org, 2016b: 1727-1736. http://proceedings.mlr.press/v48/miao16.html.

Michel P, Neubig G. MTNT: A testbed for machine translation of noisy text [C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 543-553. https://aclanthology.org/D18-1050.

Mihaylova T, Martins A F T. Scheduled sampling for transformers [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy: Association for Computational Linguistics, 2019: 351-356. https://aclanthology.org/P19-2049.

Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [C/OL]//Bengio Y, LeCun Y. 1st International Conference on Learning Representations, ICLR

2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. 2013. http://arxiv.org/abs/1301.3781.

Miyato T, Dai A M, Goodfellow I J. Adversarial training methods for semi-supervised text classification [C/OL]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. https://openreview.net/forum?id=r1X3g2_xl.

Moore R C, Lewis W D. Intelligent selection of language model training data [C/OL]//ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers. The Association for Computer Linguistics, 2010: 220-224. https://aclanthology.org/P10-2041/.

Moradi M, Samwald M. Evaluating the robustness of neural language models to input perturbations [C/OL]//Moens M, Huang X, Specia L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, 2021: 1558-1570. https://doi.org/10.18653/v1/2021.emnlp-main.117.

Och F J, Ney H. Improved statistical alignment models [C/OL]//38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 1-8, 2000. ACL, 2000: 440-447. https://aclanthology.org/P00-1056/.

Olabiyi O, Mueller E T. Multi-turn dialogue response generation with autoregressive transformer models [J/OL]. CoRR, 2019, abs/1908.01841. http://arxiv.org/abs/1908.01841.

Ott M, Edunov S, Baevski A, et al. fairseq: A fast, extensible toolkit for sequence modeling [C/OL]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 48-53. https://aclanthology.org/N19-4009.

Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation [C/OL]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 2002: 311-318. https://aclanthology.org/P02-1040/.

Park J, Sung M, Lee J, et al. Adversarial subword regularization for robust neural machine translation [C/OL]//Cohn T, He Y, Liu Y. Findings of ACL: EMNLP 2020 Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Association for Computational Linguistics, 2020a: 1945-1953. https://doi.org/10.18653/v1/2020.findings-emnlp.175.

Park J, Sung M, Lee J, et al. Adversarial subword regularization for robust neural machine translation [C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online:

Association for Computational Linguistics, 2020b: 1945-1953. https://aclanthology.org/2020.findings-emnlp.175.

Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C/OL]// Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2014: 1532-1543. https://doi.org/10.3115/v1/d14-1162.

Post M. A call for clarity in reporting BLEU scores [C/OL]//Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018: 186-191. https://aclanthology.org/W18-6319.

Press O, Wolf L. Using the output embedding to improve language models [C/OL]//Lapata M, Blunsom P, Koller A. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers. Association for Computational Linguistics, 2017: 157-163. https://doi.org/10.18653/v1/e17-2025.

Provilkov I, Emelianenko D, Voita E. BPE-dropout: Simple and effective subword regularization [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 1882-1892. https://aclanthology.org/2020.acl-main.170.

Radford A, Narasimhan K. Improving language understanding by generative pre-training [C]//2018.

Ranzato M, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks [C/OL]// Bengio Y, LeCun Y. 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. 2016. http://arxiv.org/abs/1511.06732.

Reed S E, Lee H, Anguelov D, et al. Training deep neural networks on noisy labels with bootstrapping [C/OL]//Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings. 2015. http://arxiv.org/abs/1412.6596.

Sato M, Suzuki J, Kiyono S. Effective adversarial regularization for neural machine translation [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019a: 204-210. https://doi.org/10.18653/v1/p19-1020.

Sato M, Suzuki J, Kiyono S. Effective adversarial regularization for neural machine translation [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Lin-

guistics. Florence, Italy: Association for Computational Linguistics, 2019b: 204-210. https://aclanthology.org/P19-1020.

Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016a. https://doi.org/10.18653/v1/p16-1009.

Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016b: 1715-1725. https://aclanthology.org/P16-1162.

Shao C, Feng Y, Zhang J, et al. Retrieving sequential information for non-autoregressive neural machine translation [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 3013-3024. https://doi.org/10.18653/v1/p19-1288.

Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models [C/OL]//Cortes C, Lawrence N D, Lee D D, et al. Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015: 3483-3491. https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html.

Strobelt H, Gehrmann S, Behrisch M, et al. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models [J/OL]. IEEE Trans. Vis. Comput. Graph., 2019, 25(1): 353-363. https://doi.org/10.1109/TVCG.2018.2865044.

Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C/OL]// Ghahramani Z, Welling M, Cortes C, et al. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. 2014: 3104-3112. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL]// Bengio Y, LeCun Y. 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014. http://arxiv.org/abs/1312.6199.

Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las

Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016: 2818-2826. https://doi.org/10.1109/CVPR.2016.308.

Takase S, Kiyono S. Rethinking perturbations in encoder-decoders for fast training [C/OL]// Toutanova K, Rumshisky A, Zettlemoyer L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, 2021a: 5767-5780. https://doi.org/10.18653/v1/2021.naacl-main.460.

Takase S, Kiyono S. Rethinking perturbations in encoder-decoders for fast training [C/OL]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021b: 5767-5780. https://aclanthology.org/2021.naacl-main.460.

Unanue I J, Borzeshi E Z, Esmaili N, et al. Rewe: Regressing word embeddings for regularization of neural machine translation systems [C/OL]//Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 430-436. https://doi.org/10.18653/v1/n19-1041.

Vaibhav, Singh S, Stewart C, et al. Improving robustness of machine translation with synthetic noise [C/OL]//Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 1916-1920. https://doi.org/10.18653/v1/n19-1190.

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C/OL]//Guyon I, von Luxburg U, Bengio S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 5998-6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vlasov V, Mosig J E M, Nichol A. Dialogue transformers [J/OL]. CoRR, 2019, abs/1910.00486. http://arxiv.org/abs/1910.00486.

Voita E, Serdyukov P, Sennrich R, et al. Context-aware neural machine translation learns anaphora resolution [C/OL]//Gurevych I, Miyao Y. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. Association for Computational Linguistics, 2018: 1264-1274. https://aclanthology.org/P18-1117/.

Voita E, Talbot D, Moiseev F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 5797-5808. https://doi.org/10.18653/v1/p19-1580.

Wang R, Finch A M, Utiyama M, et al. Sentence embedding for neural machine translation domain adaptation [C/OL]//Barzilay R, Kan M. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers. Association for Computational Linguistics, 2017: 560-566. https://doi.org/10.18653/v1/P17-2089.

Wei B, Wang M, Zhou H, et al. Imitation learning for non-autoregressive neural machine translation [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 1304-1312. https://doi.org/10.18653/v1/p19-1125.

Wu L, Li S, Hsieh C, et al. Stochastic shared embeddings: Data-driven regularization of embedding layers [C/OL]//Wallach H M, Larochelle H, Beygelzimer A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 24-34. https://proceedings.neurips.cc/paper/2019/hash/37693cfc748049e45d87b8c7d8b9aacd-Abstract.html.

Xie L, Wang J, Wei Z, et al. Disturblabel: Regularizing CNN on the loss layer [C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016: 4753-4762. https://doi.org/10.1109/CVPR.2016.514.

Yang M, Wang R, Chen K, et al. Sentence-level agreement for neural machine translation [C/OL]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3076-3082. https://aclanthology.org/P19-1296.

You W, Sun S, Iyyer M. Hard-coded gaussian attention for neural machine translation [C/OL]// Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020: 7689-7700. https://doi.org/10.18653/v1/2020.acl-main.687.

Yuan L, Tay F E H, Li G, et al. Revisiting knowledge distillation via label smoothing regularization [C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020: 3902-

3910. https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Revisiting_Knowledge_Distillation_via_Label_Smoothing_Regularization_CVPR_2020_paper.html.

Zhang B, Xiong D, Su J, et al. Variational neural machine translation [C/OL]//Su J, Carreras X, Duh K. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. The Association for Computational Linguistics, 2016: 521-530. https://doi.org/10.18653/v1/d16-1050.

Zhang C, Jiang P, Hou Q, et al. Delving deep into label smoothing [J/OL]. IEEE Trans. Image Process., 2021, 30: 5984-5996. https://doi.org/10.1109/TIP.2021.3089942.

Zhang J, Luan H, Sun M, et al. Improving the transformer translation model with document-level context [C/OL]//Riloff E, Chiang D, Hockenmaier J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 2018: 533-542. https://doi.org/10.18653/v1/d18-1049.

Zhang J, Zhao Y, Li H, et al. Attention with sparsity regularization for neural machine translation and summarization [J/OL]. IEEE ACM Trans. Audio Speech Lang. Process., 2019, 27(3): 507-518. https://doi.org/10.1109/TASLP.2018.2883740.

Zhang W, Feng Y, Meng F, et al. Bridging the gap between training and inference for neural machine translation [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019b: 4334-4343. https://aclanthology.org/P19-1426.

Zhang Z, Sabuncu M R. Self-distillation as instance-specific label smoothing [C/OL]// Larochelle H, Ranzato M, Hadsell R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020. https://proceedings.neurips.cc/paper/2020/hash/1731592aca5fb4d789c4119c65c10b4b-Abstract.html.

Zhao Y, Komachi M, Kajiwara T, et al. Double attention-based multimodal neural machine translation with semantic image regions [C/OL]//Forcada M L, Martins A, Moniz H, et al. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020. European Association for Machine Translation, 2020: 105-114. https://aclanthology.org/2020.eamt-1.12/.

Zhao Z, Dua D, Singh S. Generating natural adversarial examples [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=H1BLjgZCb.

Zheng B, Zheng R, Ma M, et al. Simultaneous translation with flexible policy via restricted imitation

learning [C/OL]//Korhonen A, Traum D R, Màrquez L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019a: 5816-5822. https://doi.org/10.18653/v1/p19-1582.

Zheng B, Zheng R, Ma M, et al. Simpler and faster learning of adaptive policies for simultaneous translation [C/OL]//Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, 2019b: 1349-1354. https://doi.org/10.18653/v1/D19-1137.

Zhou C, Gu J, Neubig G. Understanding knowledge distillation in non-autoregressive machine translation [C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. https://openreview.net/forum?id=BygFVAEKDH.

Zintgraf L M, Cohen T S, Adel T, et al. Visualizing deep neural network decisions: Prediction difference analysis [C/OL]//5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. https://openreview.net/forum?id=BJ5UeU9xx.

致 谢

伴随着疫情在上海的肆虐，我的学生生涯也即将迎来终点。借这个机会，我想由衷地对硕士期间帮助我的人表达感谢。

首先，我想对我的导师冯洋老师表示感恩。三年前冯老师收我为学生，对我来说是莫大的知遇之恩。硕士在读期间，冯老师从生活作风、为人处事、学术指导、学习方法、心理建设等各方面对我进行了细心的指导，督促着我不断成长，我永远不会忘记冯老师在我进步缓慢时为我担心着急的场景。除了直接指导，冯老师还以身作则，为学生树立了一个光辉的科研工作者的形象。作为一名导师，冯老师是整个课题组最刻苦的人，早出晚归，甚至通宵工作；冯老师始终坚持真正的科研精神，要求我们解决真正的问题，而不单纯为了发论文；冯老师在几年内将课题组发展壮大，建设了一流的科研环境，营造了轻松的科研氛围，给学生们提供了一个巨人的肩膀。我由衷祝愿冯老师身体健康，带领计算所自然语言处理研究组取得更辉煌的成就。

其次，我想感谢课题组的所有成员，他们是我人生的巨大财富。张文师兄、薛海洋师兄和李京谕师姐在我刚进组就毕业了，有幸见识到他们的风姿，实验室至今仍然流传着他们的传说。谷舒豪师兄是对我帮助最大的师兄，他不仅科研能力优异，还具有乐观豁达的生活态度和杰出的处事能力，为课题组的团队建设做出了重要的贡献，是我学习和生活的榜样；杨郑鑫师兄帅气潇洒，感染力强，学习上经常帮我解决问题，生活中是我的好大哥；王树根师兄工程能力非常强，很讲义气，乐于助人，我的工程任务有他一半的功劳；欧蛟师姐温柔漂亮，生活中处处关照我，经常给我零食吃；刘舒曼师姐是我的老乡，乐观随和、文静善良；申磊师姐年龄很小，活泼大方、勤勉独立；邵晨泽、单勇、李绩成师兄在学习上帮助我很多，也是我的好伙伴；李泽康和我同级，能力过人，让我无法匹敌；张绍磊、张倬诚、伍思源、马铮睿、房庆凯、刘龙祥、桂尚彤、黄浪林、赵彤钰、郭守涛、杨哲等师弟师妹成绩优异，很有礼貌，作为他们的师兄我十分汗颜。谢婉莹、董宁、赵紫毫、卫李赋凌和高博飞同学在课题组实习较长时间，很怀念和他们共处的时光。

感谢计算所的刘琳老师、程一老师、周世佳老师、敖翔老师、孟园老师、冯

刚老师和李慧老师，他们永远面带微笑地帮助我。

感谢我的室友杨树鑫同学，他很重感情，是我很好的朋友。

感谢我的父母始终不计代价地默默支持着我，感谢我的姐姐始终关心着我，希望自己能够在以后的陪伴中回报他们。

感谢在百忙之中评阅论文并提出宝贵意见的各位老师。

作者简历及攻读学位期间发表的学术论文与研究成果

基本情况:

姓名：郭登级　性别：男　出生日期：1995.01.04　籍贯：江苏连云港

教育经历:

2014 年 09 月–2018 年 06 月，于中国科学院大学物理科学学院获得理学学士学位。

2019 年 09 月–2022 年 06 月，于中国科学院计算技术研究所攻读工学硕士学位。

攻读硕士学位期间参与的工程项目:

1. 课题组机器翻译演示系统。
2. 蒙语分词和命名实体识别与翻译系统。
3. CCMT 2019 第十五届全国机器翻译大会藏汉翻译评测第一名。

攻读硕士学位期间参与的纵向课题:

1. 国家重点研发计划政府间国际科技创新合作重点专项，项目名称：基于神经网络的汉泰机器翻译研究，项目批准号 2017YFE0192900，2019/08-2022/07。
2. 国家重点研发计划科技创新 2030-“新一代人工智能”重大项目子课题，课题名称: 人机行为与情境常识的大规模知识处理与推理，课题号 2018AAA0102502，2019/12-2023/12。
3. 国家重点研发计划”前沿科技创新”子课题，子课题名称：面向多语言文本的知识抽取，课题号 2019QY2301，2019/10/31-2021/10/31。

攻读硕士学位期间发表的学术论文:

1. **Dengji Guo**, Zhengrui Ma, Min Zhang, Yang Feng. Prediction Difference Regularization against Perturbation for Neural Machine Translation, accepted by Proceedings of 60th Annual Meeting of the Association for Computational Linguis-

tics, main conference, long paper.

2. Yang Feng, Shuhao Gu, **Dengji Guo**, Zhengxin Yang, and Chenze Shao. 2021. Guiding teacher forcing with seer forcing for neural machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(ACL/IJCNLP 2021), (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 2862–2872. Association for Computational Linguistics.
3. 谷舒豪, 单勇, 谢婉莹, **郭登级**, 王树根, 邵晨泽, 薛海洋, 张良, 冯洋. 基于数据增强及领域适应的神经机器翻译技术. 江西师范大学学报 (自然科学版), 2019 年 11 月, 第 43 卷, 第 6 期.