Computer Science and Application 计算机科学与应用, 2020, 10(3), 387-397
Published Online March 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.103040

Hans汉斯

Analyzing the Problems of Vocabulary in Japanese-Chinese Neural Network Machine Translation

Wentao Luo

Kodensha, Acronyms Acceptable, Osaka Japan
Email: rabuntou@kodensha.jp

Received: Feb. 6th, 2020; accepted: Feb. 21st, 2020; published: Feb. 28th, 2020

Abstract

In recent years, Neural Network Machine Translation (NMT) has made great progress as a new translation technology. Its translation results are not only more accurate but also more fluid. But at the same time, NMT also has many problems that need to be solved. The purpose of this article is to explore problems of vocabulary and their causes, and propose solutions for tuning model of Japanese-Chinese NMT. The limitation of the size of vocabulary and the domain mismatch of corpus could lead some problems such as unknown words and mistranslated words. Therefore, this article proposes several solutions like using subword, replacing low-frequency words, using external dictionaries, and using domain adaptation. Using subword or using external dictionary can overcome the problem caused by small size of vocabulary. Replacing low-frequency words can reduce the negative influence of low-frequency words. Domain adaptation can improve the performance on translating specific domain text. The experimental results showed that compared with the general NMT model, the approaches of tuning model proposed in this article can reduce the number of vocabulary translation problems and improve the translation quality.

Keywords

Neural Network Machine Translation, Vocabulary Problems, Tuning Model

关于日中神经网络机器翻译中的词汇问题的探讨

罗雯涛

株式会社高电社，日本 大阪
Email: rabuntou@kodensha.jp

收稿日期：2020年2月6日；录用日期：2020年2月21日；发布日期：2020年2月28日

文章引用: 罗雯涛. 关于日中神经网络机器翻译中的词汇问题的探讨[J]. 计算机科学与应用, 2020, 10(3): 387-397.
DOI: 10.12677/csa.2020.103040

摘　要

近年以来，神经网络机器翻译作为新兴的翻译技术，取得了极大的进步。翻译的译文不仅更加准确也更为流畅。但神经网络翻译同时还有许多问题需要改进。本文旨在以日中神经网络机器翻译为实例，探讨词汇层面的问题和成因，并提出相应的模型改进方法。受限于模型的词表大小和语料资源的领域不匹配等原因，译文中存在未知词和词语的错翻漏翻等问题。因此，本文根据这些原因提出了使用subword，替换低频词，利用外部词典，采用领域自适应训练模型等多个改进方案。使用subword或者利用外部词典，可以克服词表过小的问题。替换低频词可以降低低频词对模型的负影响。领域自适应可以提高模型对特定领域文本的表现。实验结果表明本文提出的模型改进方案相较于一般的神经网络翻译模型，能很好地减少词汇翻译问题的出现次数，从而提高译文的翻译质量。

关键词

神经网络机器翻译，词汇问题，模型改进

Copyright © 2020 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

机器翻译指的是通过计算机自动地将一种语言翻译成另一种语言的技术。它的作用是消除人们的沟通障碍，是人类长期以来的一个钻研方向。

本文的目的是通过探讨日中神经网络机器翻译存在的词汇问题，了解这些问题产生的原因，并实施若干的改进方法。最后通过分析实验的结果，验证这些方法是否能有效改进词汇问题。

2. 机器翻译的类型

在介绍神经网络机器翻译存在的词汇问题之前，本文先简要总结机器翻译的各种类型。根据出现时间的先后顺序，机器翻译主要可以分为以下几种类型：基于规则的机器翻译，基于实例的机器翻译，基于统计的机器翻译，基于神经网络的机器翻译。

2.1. 基于规则的机器翻译

基于规则的机器翻译主要是受到了乔姆斯基提出的转换生成语法的启发。乔姆斯基[1]认为对于一种语言，是可以利用有限的规则来推导出来无限的句子。

基于规则的机器翻译的优点在于他可以直观并精确地表达语言学家们所制定的各种知识规则。同时，其缺点也是非常明显的。比如规则需要动用大量的人力来进行编写；且规则具有极大的主观性，难以保障一致性；不规范的句子难以被归纳到有限的规则里。

2.2. 基于实例的机器翻译

基于实例的机器翻译的方法最早由日本京都大学的长尾真[2]提出。其思想是要给机器翻译系统提供已经存在的近似的例句，使其在翻译新输入的句子时可以忽略其中的相同部分，可以专注于处理例句和新输入的句子之间的不同的部分。

比如要翻译“我要去电影院”这样一句话时，若翻译系统知晓已知的例子里有“我要去剧院”，那么只需要考虑如何翻译和处理“电影院”即可。这种方法不要求花大量时间来定义规则，并在之后一定程度上促进了之后的统计机器翻译。

2.3. 基于统计的机器翻译

随着数学理论的介入，机器翻译开始以统计模型作为基础。统计机器翻译的基本思想是对大量的平行语料进行统计分析，以此来构建统计翻译模型，并在这模型的基础上定义要估计的模型参数，并设计参数估计算法。统计机器翻译中的一个代表是基于短语的机器翻译[3]。

2.4. 基于神经网络的机器翻译

近年来，随着深度学习的快速发展，神经网络的架构也被引入了机器翻译的研究中。基于神经网络的机器翻译(Neural Machine Translation, NMT) [4] [5]也逐渐兴起。相较于统计机器翻译，NMT 需要使用更大的语料，能够得到更精确的译文和更流畅的语句。

针对此，学术界和产业界也投入大量的人力物力开发各自的神经网络机器翻译系统，并取得了巨大的进步。在许多语种上，神经网络机器翻译的性能都得到了大幅提升，并远远超越了传统的统计机器翻译。

从使用的模型框架的角度来对 NMT 进行分类的话，目前的 NMT 模型主要包含三类：基于循环神经网络(Recurrent Neural Network, RNN)的 NMT 模型，基于卷积神经网络(Convolution Neural Network, Conv)的 NMT 模型，基于自注意力机制的 NMT 模型(Transformer)。

3. 神经网络机器翻译中存在的词汇问题

神经网络机器翻译的翻译质量比起传统的机器翻译方式在翻译质量上得到了明显提高，译文更加准确和流畅。但神经网络机器翻译的研究中仍存在许多问题需要解决。其中，一个最重要问题是词汇相关的问题。

可以想象到如下一个应用场景：专业文本中往往存在着大量的专业词汇，比如法律政令，科技论文等等。这些词汇具有一定的翻译难度。当用户需要对专业领域性质较强的文本时，由于神经网络机器翻译已经将句子翻译得很流畅，其往往也会更在意词汇有没有被正确翻译。

因此词汇问题是神经网络机器翻译急需解决的一个研究方向。

3.1. 词汇问题

从在译文所呈现的情形来看，词汇问题可以分为“未知词”和“翻译错误”这两种情况。本文以实际日中神经网络机器翻译的译文为例，描述这些问题。

3.1.1. 未知词

未知词是指在翻译过程中，原文部分的单词未经翻译，直接出现在了译文中的词。如下面这个例句：

(例句 1)原文：庶民寄りの政策趣旨がいくら好ましくても、国の財政を湯水のように使う「免罪符」となってはならない。

(例句 1)译文：不管怎样出色的支持庶民政策，都不应将其用作使用政府资金如热水的“免罪符”。

当神经网络机器翻译系统不能判断该输出哪个译文时，系统会倾向于从原文中寻找一个概率最大的原文单词进行替代。例句 1 中，日文的“庶民”被直接输出到了中文译文里。这里的“庶民”应该被翻译为“老百姓”“大众”。

(例句 2)原文：現在は一般的なノートパソコンのポインティングデバイスとして広く採用されている。

(例句 2)译文：如今，它被广泛用作普通笔记本计算机的ポインティングデバイス。

又如例句 2 中，日文的专业名词“ポインティングデバイス”没有被正确地翻译为中文。而“ポインティングデバイス”的正确译文应该是 “定点设备”“指向设备”。

3.1.2. 错误翻译

错误翻译，即系统错误地翻译了该单词，甚至漏翻了该单词。

(例句 3)原文：協会は、新たに八百長容疑が明らかになった 14 人の力士に対しても、独自調査を進めている。

(例句 3)译文：该协会正在对 800 人的罪名的 14 名相扑选手进行独立调查。

在例句 3 中，日文单词“八百長”是一个特殊的词汇，其意思为“体育比赛当中的故意让赛的行为”。译文错误地把“八百長”当作一个“数词+量词”的短语进行了翻译。于是译文变成了“800 人”。

(例句 4)原文：移動通信の重要な基盤技術の 1 つである狭帯域デジタル変調方式について概要を述べる。

(例句 4)译文：本文概述了数字调制方案，这是移动通信的重要基础技术之一。

在例句 4 中，日文单词“狭帯域デジタル変調方式”中的“狭帯域”并没有被翻译出来，其对应的中文应当是“窄带(宽)”。本例中可以看到，在翻译一些复杂的复合词时，神经网络机器翻译有可能会出现错翻漏翻。

3.2. 问题产生的原因

从问题的原因入手，词汇问题又可以被划为如下两种情形。

3.2.1. 受词表限制

词表限制是指因为过大的词表会造成计算开销过大的问题，神经网络机器翻译的模型一般使用固定大小的词表。

于是在训练翻译模型的过程中，对于源语言，模型并未能很好地学习到词汇表以外的词汇的语义表示。对于目标语言，模型并不能选择词汇表以外的词汇作为译文。前者很可能会导致错误翻译，后者则很可能导致出现未知词。

因为词表大小的限制而排除在词表外的词又称作未登录词(out of vocabulary)。未登录词的存在将极大地影响翻译的质量。因此如何解决词表受限的问题成为改善词汇问题的一个重点。

根据 Sennrich 等人[6]的报告，不光是未登录词，系统在翻译这些存在于词表中的低频词的时候，其翻译质量也不尽如人意。所谓低频词指的是这些词存在于 NMT 的词表中，但由于在训练语料中出现的频次数过少。由于出现次数太少，使得这些词的语义表现也并没有被翻译模型充分吸收。

3.2.2. 受资源限制

神经网络机器翻译的模型的训练需要使用到大量的语料。语料的不足一般被认为会导致训练后的模型的翻译质量较差。这个情形在翻译特定领域的文本时，表现地尤为明显。这个结果在很大程度上归结于“领域不匹配”。由于收集特定领域的语料有一定难度，使得训练模型所使用的语料中该领域所占比可能过小，训练出的翻译模型对于特定测试集出现领域不匹配的问题。

Koehn 的报告[7]阐明了领域不匹配时，翻译测试结果将会降低。如表 1 所示，该文给出了使用不同语料训练的翻译模型(纵轴)对不同领域的测试集(横轴)的翻译测试结果(表中的结果值为自动评价方法 BLEU)。

Table 1. Performance of the NMT models trained from different corpora on different test sets
表 1. 不同语料训练得到的 NMT 模型在不同领域的测试集上的表现

Model	LAW	Medical	IT	Subtitles
All data	30.5	45.1	35.3	26.4
LAW	31.1	12.1	3.5	2.8
Medical	3.9	39.4	2.0	1.4
IT	1.9	6.5	42.1	3,9
Subtitles	7.0	9.3	9.2	25.9

可以发现，使用 ALL data (包含 LAW, Medical, IT, Subtitles)作为训练语料的翻译模型在各个测试集上的表现有着很大的波动。比如翻译 Medical 领域时有 45.1 的 BLEU 值，而翻译 Subtitles 领域时仅有 26.4 的 BLEU 值。 这说明了一些特定的领域的文本，本来就有一定的翻译难度。若翻译模型的训练语料缺少该领域的语料时，翻译质量肯定会降低。

而通过考察训练语料没有相关的领域的语料时的情形，可以进一步看到其翻译表现的降低程度是不容忽视的。比如，使用 ALL data 训练的系统在翻译 Medical 领域时有 45.1 的 BLEU 值，但仅使用 LAW 语料进行训练的模型在翻译 Medical 领域时只有 12.1 的 BLEU 值。

4. 模型改进方法

针对词汇问题的成因，本文提出以下几种改进模型翻译效果的方案。

4.1. 解决词表受限的方法

为了从词汇表受限的角度解决词汇问题，本文采用了以下三种方法。

4.1.1. 使用 subword 代替传统单词

Subword 指代的是一种比单词颗粒度更小的语言单位。Sennrich 等人[6]为了解决词汇表受限的问题，提出了切割出一种比单词更小的语言单位作为词表中的“单词”来使用的方法。从而使得这些“单词”可以被词表全数收录。本文在分割 subword 时，参考和利用了工具 SentencePiece [8]。

由于中文或日文等亚洲语言没有天然的单词界限，因此往往需要先进行分词处理。Sentencepiece 对此问题进行了改良，使得 subword 的分割可以不需要事先的分词处理。使用工具 Sentencepiece 实际分割后的例句如下：

(例句 5)日文原文：スクリプトの開発に開発エンドポイントを使用する

(例句 5)日文 subword：_スクリプト_の_開発_に_開発_エンドポイント_を使用_する

(例句 5)中文原文：使用开发终端节点来开发脚本

(例句 5)中文 subword：_使用_开发_终端节点_来开发_脚本

例句中的下划线表示了此处是实际单词的词头。一般而言，使用分词工具切分时，日文的“エンドポイント”是一个单词，中文的“终端节点”也是一个单词。分割 subword 会将语言切成了颗粒度更小的单位，在本例中，上述单词分别被切分为了 2 个和 3 个颗粒度更小的单位。

与此同时，对于训练翻译模型来说，使用分割成 subword 的语料进行训练，可以学习到更丰富的语言表现。

4.1.2. 替换低频词

替换低频词的思想是，因为低频词也并不能被模型很好地学习到，那么若能在训练模型之前，将训

练语料中的低频词替换为意思相近的高频词，则可以改善最终的模型的翻译质量。本文认为，语料中出现频次低于 3 的单词，是会对翻译模型产生负影响的低频词。

关于词的替换，我们将以词的相似度为替换标准。具有较高相似度的两个词，可以被判断为可替换词。

本文使用 Gensim 工具库[9]来计算词向量和词的相似度。具体的步骤如下。

首先，需要获取单词的词向量。本文利用 Gensim 工具库，得到基于 Word2Vec [10] [11]的词向量空间模型。

本文选择使用维基百科的 dump 数据作为获取词向量模型的语料，维基百科的 dump 数据可从以下地址获取(https://dumps.wikimedia.org/)。

第二步，得到词向量之后，通过工具库提供的 similarity 函数，可以计算两个词之间的相似度。相似度值的范围在 0 到 1 之间。

第三步，进行词的替换。对于语料中具体的句子，若其中的单词为低频词或者是未登陆词，将通过计算语义相似度的方式，从高频词中寻找最为匹配的可替代的词。若匹配成功，则进行替换。工具库提供的 most_similar 函数，可以有效的节省匹配时间。

最后还需要处理面对多个可替换候补如何进行选择的问题。在实际的替换过程中，本文参照以下原则进行依次优先选择。

a. 比较各个候补词在训练语料中的词频大小差异，若词频大小有明显差距，则选取词频更高的候补词。根据实际语料分析，词频的明显差距定为 10 次。

b. 比较各个候补的相似度差异大小，若语义相似度差距明显，则选取相似度更高的候补词。根据实际语料分析，相似度的明显差距定为 0.2。

c.选择词频最高的候补词。

(例句 6)替换前：此事需要考虑到当地的具体情况，不能中央凭空拍板，应责成当地地方领导亲自督办。

(例句 6)替换后：此事需要考虑到当地的具体情况，不能中央凭空拍板，应安排当地地方领导亲自督办。

在例句 6，通过词的替换，低频词“责成”被替换为更常见的高频词“安排”。

4.1.3. 外部词典

虽然翻译模型的词表大小是有限的，但独立于模型之外的外部词典则可以满足无限的词汇量需求。翻译系统可以利用外部词典，把未知词处理成相应的译文。

而关于如何获取外部词典的资源的问题，有两种可以考虑的方法。

一种是通过维基百科获得词条及其对应的译文词条作为外部词典。这类数据以词对为单位。基于维基百科获取词条的具体步骤如图 1 所示：

该流程的一开始将选择维基数据的某个分类(category，对应于某个专业领域)的页面作为起始页面，递归地向下探索其子分类的页面。当探索达到足够探索深度时终止之。针对探索过的页面，获取页面对应的词条的同时，并查询页面对应译文页面。若存在译文页面，则获取译文页面的词条。最终将词条和译文词条成对地存储到词典里。

通过这样的方法，可以快速简单地集成一个具体领域的专业词典，作为翻译模型的外部词典使用。这对于解决翻译专业领域文本时大量出现的未知词问题，很有帮助。

另一种获取词典的方法是通过基于统计的翻译方法，获得短语翻译表(phrase-table)，然后从短语翻译表中抽取出适当的短语对存入词典。

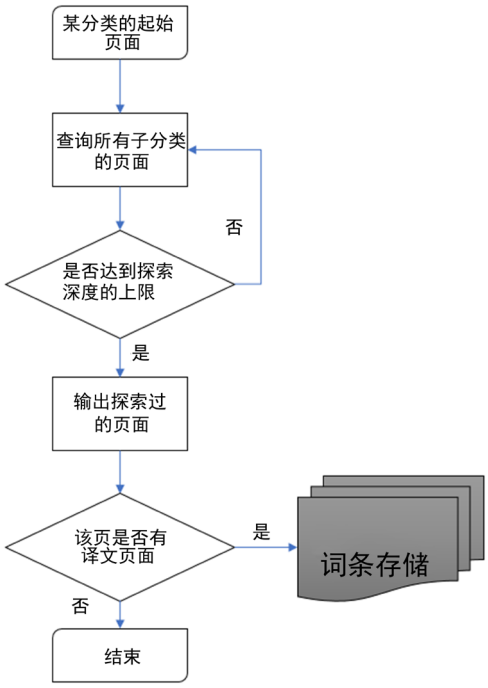

Figure 1. Flowchart: Extracting word pairs from Wikidata
图 1. 流程图：从维基的数据中抽取词条

在基于统计的翻译方法中，短语翻译表的一条短语对的记录可以表示为如下例子。

(短语翻译表的例子)アンタゴニストの同定方法|||阻断剂的认定方法|||0.75 1.22362e−07 1 0.0273384 ||| 0-0 1-1 2-2 3-3

上述记录中，原文单语，译文短语，翻译概率，词对齐信息等四项信息分别被“|||”符号分割开来。这里的第三项信息“翻译概率”中包含了基于统计的翻译过程中所感知到的四种翻译概率。

根据训练语料的规模不同，短语翻译表所记录的短语对的数量可以从数百万条到数千万条不等。但并不是每一条记录都是对于收集外部词典而言是有用的。因而，本文以翻译概率为依据，从短语翻译表中抽取出翻译概率较高的短语对，作为外部词典使用。

通过计算的各个翻译概率 P 之间的 Cross Entropy (式 1)，获得此词条的权重值 W。根据真实语料分析，当权重 W 大于−0.5 时，其记录作为词条的准确性是可以值得信赖。

$$W = \frac{1}{N}\sum_{N} \log Pn \tag{1}$$

4.2. 解决资源受限的方法

资源受限的最大表现就是语料的领域不匹配。解决领域不匹配的最直接的方法就是追加相应领域的语料。

由于特定领域的语料获得难度很大，往往只能获得一个小规模的数据。在这样的背景下，仅仅简单地追加对应领域的训练语料，由于其在训练语料中的比重和规模都较小，并不能直接改善翻译表现。因此还需要利用领域自适应(Domain adaptation)的方法强化翻译模型在特定领域上的表现。

领域自适应的目标领域又可称为“内领域”，其语料获取困难，通常语料规模较小。于此相对的，广泛存在的，易于获取语料的部分则称为“外领域”。领域自适应的训练方法可以帮助模型尽可能有效地学习到内领域语料的知识表现。

关于领域自适应的方法，本文参考了 mixed fine tuning 的方法[12] [13]，将翻译模型的训练分为三次进行。

如图 2 所示，在第一次的训练中，仅利用外领域的语料进行训练，获得了外领域模型。第二次训练以外领域模型为基础，将内外领域的语料进行混合训练，获得了交叉模型。第三次训练，以交叉模型为基础，通过训练内领域的语料，获得了最终的内领域模型。

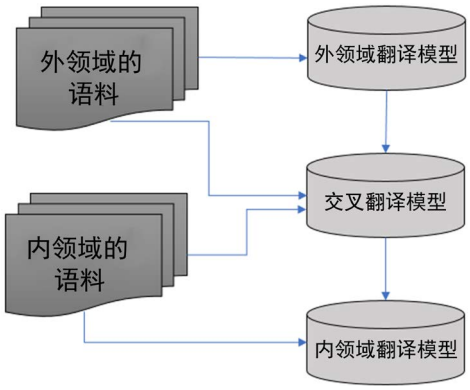

Figure 2. Domain adaptation: the approach of mixed fine tuning
图 2. 领域自适应：mixed fine tuning 方法

特别需要注意的是在训练交叉模型时，要根据内外领域语料的数量差异，对数量较少的内领域语料做增量处理(oversampling)，使之规模与外领域的语料持平。

5. 实验结果

5.1. 训练语料及测试集

本文采用 OPUS [14]和 ASPEC [15]的日中对齐语料作为训练语料，共计约 160 万句。测试集分为两个。一个为 1000 句不属于特定领域的句子集，即“普通测试集”。另一个为 1000 句的 IT 领域句子集，即“IT 测试集”。测试集和训练用的语料之间不存在交集。

依照 4.1 节描述的方法，在采用低频词替换的方法时，本文总计从语料中替换了 1 万 8 千余处低频词。针对使用外部词典的方法，本文从 wiki 数据中抽取了 1 万 3 千余条词条作为外部词典。从单词翻译表中抽取了 2 万余条短语对作为外部词典。

根据 4.2 节描述的方法，另从 OPUS 的语料库中抽取了和 IT 领域句子相近的 10 万条句子，作为领域自适应中追加训练用的 IT 领域的语料。

5.2. 实验设置

本文训练神经网络翻译模型的工具为 OpenNMT-py [16] (v0.8)。模型采用基于循环网络结构(RNN)的编码器-解码器架构。模型的词表大小设定为 5 万词。根据采用的改进方法的不同，本文将分别设置了多个评测任务。

Baseline：作为基本评测任务，代表了模型没有运用本文中提到的任何改进方法。

+Subword：表示此任务在训练模型时，对训练语料进行了 subword 分割。

+Replace：表示此任务对语料中的低频词进行了替换处理。

+Dic：表示此任务中会利用外部词典。

+Adaptation： 表示此任务针对 IT 领域进行了领域自适应的模型训练。

5.3. 实验方法

关于结果的评测方法，本文采用了自动评价和人工评价两种方法。

自动评价采用了 BLEU 值[17]作为评价标准。自动评价的目的是比较经过方法改进后，译文整体质量的变化。而人工评价则主要是人工地统计译文中存在的词汇层面的翻译问题的数量。人工评价的目的是考察经过方法的改进，译文中的词汇问题的减少情况。

Table 2. The results of automatic evaluation (BLEU value)
表 2. 自动评测结果(BLEU 值)

任务名	普通测试集	IT 测试集
Baseline	30.28	22.16
+Subword	31.66	23.48
+Replace	30.45	23.03
+Dic	30.87	24.29
+Adaptation	29.95	24.10
+Subword+Adaptation	30.15	24.32
+Replace+Dic+Adaptation	30.42	24.46

表 2 记录了自动评测的结果。作为最基本的任务，Baseline 在普通测试集的 BLEU 值为 30.28，在 IT 测试集上的 BLEU 值为 22.16。Baseline 在普通测试集上的表现要优于在 IT 测试集上的表现。这也证明了在模型的训练语料的领域不匹配时(对于 IT 领域的测试集)，模型的翻译质量会下降。

对于普通测试集而言，以解决词汇表受限为目的的三个方法的相关任务相较于 Baseline，BLEU 值都获得了一定的提升。其中最为有效的任务+Subword 的 BLEU 值为 31.66。而在使用领域自适应的方法之后，+Adaptation 的任务在普通测试集上的翻译效果反而会存在些许下降。+Adaptation 的 BLEU 值为 29.95。

另一方面，对于 IT 测试集而言，基于各种改进方法的任务在这个测试集上均相对于 Baseline 获得了提升。这证明了本文所提议的改进方法的有效性。其中以解决词汇表受限为目的的三个方法里，利用外部词典的+Dic 带来的提升最大，其 BLEU 值为 24.29。领域自适应(+Adaptation)的方法也同样获得了较大的提升，BLEU 值为 24.10。在混用多个方法的任务中，+Replace+Dic+Adaptation 任务达成了所有模型中的最高 BLEU 值 24.46。

Table 3. The results of human evaluation (Number and reduction rate of vocabulary problems)
表 3. 人工评测结果(词汇问题的数量/词汇问题的降低率)

任务名	普通测试集	IT 测试集
Baseline	127	311
+Subword	5 (96%)	6 (98%)
+Replace	109 (15%)	237 (24%)
+Dic	76 (40%)	106 (66%)
+Adaptation	125 (1%)	151 (53%)
+Subword+Adaptation	9 (93%)	5 (98%)
+Replace+Dic+Adaptation	82 (35%)	53 (83%)

在人工评测中，当出现未知词或单词的错翻漏翻等现象时，其将被计为 1 处词汇问题。表 3 记录了人工评测的结果，括号外的数值为词汇问题的数量，括号内的数值为相较于 Baseline，词汇问题减少的百分比。

Baseline 在普通测试集上有 127 处词汇问题，在 IT 测试集上的词汇问题更多，为 311 处。Baseline 在普通测试集上的表现要优于在IT测试集上的表现。这种情形在前面的自动评价的结果中也得到了印证，说明了在训练语料的领域并不匹配时，模型的翻译质量会下降。

在于普通测试集而言， 使用领域自适应(+Adaptation)的方法并未能有效减少词汇问题的数量，其减少率仅为 1%。这也从侧面印证了为何自动评测中+Adaptation 的 BLEU 值会比起 Baseline 还稍有降低。而+Subword 在减少词汇问题数量上的表现非常显著，达到了 96%。译文中几乎不存在未知词，仅有个别词语被错翻漏翻。其次，+Dic 的方法也极大地改善了词汇问题，其词汇问题减少率为 40%。因此可以认为，+Subword，+Replace，+Dic 所代表的改进方法对于改进普通测试集的译文是有效的。

另一方面，对于 IT 测试集而言，使用领域自适应(+Adaptation)的方法开始表现出作用，其词汇问题减少率达到了 53%。根据结果，可以看到+Subword, +Dic, +Dic+Adaptation 三种测试任务所代表的改进方法对于改进 IT 测试集的译文是有效的。同时，在这个测试集上，综合多种方法的+Replace+Dic+Adaptation 的任务也能够达成 83%的词汇问题减少率。

6. 总结

本文以日中神经网络机器翻译为例，探讨了神经网络机器翻译中可能遇到的词汇方面的问题，分析了问题产生的原因，并提出了若干个改进问题的方法。最后通过进行实验并分析实验结果，确认了这些方法的有效性。

参考文献

[1] Noam, C. (1956) Three Models for the Description of Language. *IRE Transactions on Information Theory*, **2**, 113-124. https://doi.org/10.1109/TIT.1956.1056813

[2] Nagao, M. (1984) A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In: *Artificial and Human Intelligence*, Elsevier Science Publishers, New York.

[3] Koehn, P., Och, F.J. and Marcu, D. (2003) Statistical Phrase-Based Translation. *Proceedings of the* 2003 *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127-133. https://doi.org/10.21236/ADA461156

[4] Bahdanau, D., Cho, K. and Bengio, Y. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. In: *Proceedings of the* 3*rd International Conference on Learning Representations* (*ICLR* 2015), International Conference on Learning Representations, San Diego, CA.

[5] Luong, T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. In: *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Lisbon, Portugal, 1412-1421. https://doi.org/10.18653/v1/D15-1166

[6] Sennrich, R., Haddow, B. and Birch, A. (2016) Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, ACL, Berlin, Germany, 1715-1725. https://doi.org/10.18653/v1/P16-1162

[7] Koehn, P. and Knowles, R. (2017) Six Challenges for Neural Machine Translation. In: *Proceedings of the First Workshop on Neural Machine Translation*, Association for Computational Linguistics, Vancouver, 28-39. https://doi.org/10.18653/v1/W17-3204

[8] Kudo, T. and Richardson, J. (2018) SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. CoRR. https://doi.org/10.18653/v1/D18-2012

[9] Radim, Ř. and Sojka, P. (2010) Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of LREC* 2010 *Workshop New Challenges for NLP Frameworks*, University of Malta, Valletta, Malta, 46-50.

[10] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space.

ICLR.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. NIPS.

[12] Chu, C.H., Dabre, R. and Kurohashi, S. (2017) An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In: *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-2061

[13] Chu, C.H., Dabre, R. and Kurohashi, S. (2018) A Comprehensive Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. *Journal of Information Processing*, **26**, 1-10. https://doi.org/10.2197/ipsjjip.26.529

[14] Tiedemann, J. (2016) OPUS-Parallel Corpora for Everyone. *Baltic Journal of Modern Computing* (*BJMC*), *Special Issue*: *Proceedings of the* 19*th Annual Conference of the European Association of Machine Translation* (*EAMT*), **4**.

[15] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H. (2016) ASPEC: Asian Scientific Paper Excerpt Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2016).

[16] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A.M. (2017) OpenNMT: Open-Source Toolkit for Neural Machine Translation. CoRR. https://doi.org/10.18653/v1/P17-4012

[17] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the* 40*th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002, 311-318. https://doi.org/10.3115/1073083.1073135