华光机器翻译系统 CCMT2022 评测技术报告

殷建民*，郑文亮，王淑珍

（潍坊北大青鸟华光照排有限公司，山东潍坊 261205）

摘 要： 本文描述了华光机器翻译系统参加 2022 年全国机器翻译大会（CCMT 2022）翻译评测任务的情况。本次评测中，我们参与了蒙汉综合领域、藏汉综合领域和泰中综合领域的 3 个评测项目，提交了测试集翻译结果。本文对评测所用系统、技术路线以及系统运行软硬件环境进行了详细的介绍。

关键词： 机器翻译；评测；蒙古文；藏文；泰文

The CCMT2022 evaluation technology report of Huaguang Machine Translation Systems

Jianmin Yin, Wenliang Zhang, Shuzhen Wang

(Weifang Beida Jade Bird Huaguang Information Technology Co., LTD, Weifang 261205）

Abstract: This article describes how machine translation system participated in the translation evaluation task of the 2022 National Conference on Machine Translation (CCMT 2022). In this evaluation, we participated in 3 evaluation projects in the Mongolian-Chinese comprehensive field, the Tibetan-Chinese comprehensive field, and the Thai-Chinese comprehensive field, and submitted the translation results of the test set. This paper introduces the system used in the evaluation, the technical route and the hardware and software environment of the system in detail, and analyzes the evaluation results.

Key words: machine translation; evaluation; Mongolian; Tibetan; Thai

1 引言

本届机器翻译评测，本公司参加了 3 个项目：蒙汉综合领域机器翻译（MC）、藏汉综合领域机器翻译（TC）和泰中低资源语言机器翻译（ThaiC）。本次评测所提交的系统使用近年来比较成功的神经网络机器翻译（NMT）模型作为核心。

本文将主要介绍我们采用的机器翻译系统框架，主要技术以及在 3 个评测项目中的性能表现。

2 项目背景与研发过程

本公司参评的蒙汉、藏汉机器翻译系统均为“国家数字复合出版系统工程--少数民族文字出版资源管理系统及辅助工具”的子课题。泰中机器翻译系统则借鉴了“国家数字复合出

基金项目： 国家新闻出版广电总局国家新闻出版重大科技工程专项资金项目（XWCB-GDGC-FHCB/16）。

* **通信作者：** 110154413@qq.com。

版系统工程--少数民族文字出版资源管理系统及辅助工具”的“傣汉机器翻译系统”。

2014 年 12 月 10 日，本公司与国家新闻出版广电总局签订项目合同书。2021 年 12 月 29 日，本项目通过终验评审。

3 系统

系统采用了基于神经网络的机器翻译(Neural Machine Translation, NMT)技术和编码器-解码器(encoder-decoder)框架，将任意长度的输入句子首先编码成为一个固定长度的不可解释的向量，然后根据这个向量进行解码自动生成译文。其最大优势是可以使用户根据自身需求任意修改系统，同时绝对保障用户翻译数据安全。所有的翻译数据，包括原始数据、翻译结果、翻译过程中产生的日志信息等，都将保存在自己的服务器上，能确保以安全的方式完成翻译任务。

3.1 系统框架

系统采用 NMT 的 encoder-decoder 模型，encoder 把源语言序列进行编码，并提取源语言中信息，通过 decoder 再把这种信息转换到另一种语言即目标语言中来，从而完成对语言的翻译。

对于提供的双语语料，我们进行训练翻译规则和 NMT 模型。对于单语语料，我们进行训练前向语言模型以及后向语言模型。所有模型均作为对数线性模型的一个特征融入到解码器中，从而通过翻译源端句子得到目标译文。整个系统的框架如图 1 所示。

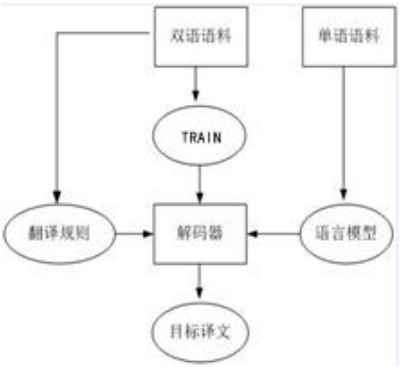

图 1 系统框架

Fig.1 System framework

3.2 编码器

本系统采用带注意力机制的 seq2seq 学习[1]，这种机制的主要作用就是在预测一个目标词汇的时候，它会自动地查找源语言序列中哪一部分与之相对应，并且在后续的查找生词中可以直接复制相对应的源语言词。

本系统 encoder 采用双向 RNN（bi-directional RNN），由 RNN 有前向和后向 RNN 组成。前向 RNN$\overrightarrow{f}$正向读取输入序列（从x_1到x_T），并计算前向隐藏层状态$(\overrightarrow{h_1}, \ldots, \overrightarrow{h_T})$,而后向 RNN$\overleftarrow{f}$从反向读取输入序列（从$x_T$到$x_1$），并计算反向隐藏状态$(\overleftarrow{h_1}, \ldots, \overleftarrow{h_T})$。对于每个单词$x_j$，我们把它对应的前向隐藏状态向量$\overrightarrow{h_j}$和后向隐藏状态向量$\overleftarrow{h_j}$拼接起来来表示对$x_j$的注解

（annotation），例如$h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}]$，这样，注解$h_j$就包含了所有词的信息。由于 RNN 对最近的输入表达较好，所以注解h_j主要反映了x_j周围的信息[2]。

3.3 解码器

在解码的结构中，定义条件概率：

$$p_{(y)} = \prod_{t=1}^{T'} p(y_t|\{y_1, \ldots, y_{t-1}\}, c)$$

$$p(p(y_t|\{y_1, \ldots, y_{t-1}\}, c)) = g(y_{t-1}, s_t, c)$$

其中，g为非线性函数，s_t是 decoder 的隐藏状态，c 是由 encoder 的隐藏序列产生的上下文向量，这个具体是什么等一会说。

把上式的条件概率写为：

$$p(y_i|y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i)$$

其中，s_i是时间步i的隐藏状态，可由下式来计算：

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

上下文向量c_i依赖于一系列的注解$(h_i, \cdots, h_t)$，上下文向量是由这些注解h_j加权求和算出来的：

$$c_i = \sum_{j=1}^{T} a_{ij} h_j$$

每个注解h_j的权重a_{ij}由下式计算：

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{ik})}$$

其中，$e_{ij} = a(s_{i-1}, h_j)$为对位模型(alignment model)，它是计算位置j周围的输入与位置i的输出相匹配的得分函数。而向量$a_i = (a_{i1}, a_{i2}, a_{iT})$为注意力向量，又为词对位向量。

3.4 注意力机制的改进

注意力机制分为全局注意力(Globale attention)机制和局部注意力(Local attention)机制，简单的来说，是使用全部的 encoder 的隐藏层还是部分。Global attention 都是使用了全部的 encoder 的隐藏状态。Local attention 就是选择一个较小的上下文向量窗口。

首先还是要获得这个上下文向量c_t，这个上下文向量用来捕获源语言的相关信息来预测目标词y_t，然后把 decoder 的隐藏状态s_t与这个上下文向量c_t拼接起来通过非线性函数产生注意力隐藏状态(attentional hidden state)：

$$\overline{h_t} = tanh(W_c[c_t; s_t])$$

最后，使用 softmax 函数进行预测：

$$p(y_t|y_{<t}, x) = softmax(W_s \overline{h_t})$$

4 评测过程和方法

4.1 评测数据准备

本次我们参与的机器翻译评测语料包含 3 种语言：蒙汉、藏汉和泰中，均为综合领域。

在评测中输入输出文件均采用 UTF-8 编码（有 BOM）以及严格的 XML 格式。

本次使用的训练集、开发集、测试集均为评测方提供的训练数据。具体评测数据情况如表 1 所示：

表 1　CCMT 2022 评测翻译模型训练集数据情况

Tab.1 CCMT 2022 evaluation of translation model training set data

评测项目名称	**训练数据**（双语）	**开发数据**（句）	**测试数据**（句）
蒙汉综合领域机器翻译	126万	10000	10000
藏汉综合领域机器翻译	115万	10000	10000
泰中综合领域机器翻译	19万	5000	5000

5 技术路线

5.1 语料预处理

我们对所有语料（双语训练语料，开发集，测试集，语言模型训练语料）进行了一定的预处理，下面我们对这四种语料预处理来做一些详细介绍。

蒙古文：

具体蒙古文语料预处理如下：

- 行尾转换
- 全角转半角
- 转义字符的处理
- 词干切分
- 短语处理
- 领域分类
- 过滤一些包含错误编码的句子
- Tokenisation 处理
- 标点与符号的处理
- 分隔符规范化处理

藏文：

采用西北民族大学李亚超提供的藏文分词工具进行分词处理。构建短语表时，融合了的单词翻译概率，以避免某些单词的翻译不在短语表中的情况。根据藏文语言特点处理藏文语序和汉语差异的情况。具体藏文语料预处理如下：

- 行尾转换
- 全角转半角
- 转义字符的处理
- 分词
- 领域分类
- Tokenisation 处理
- 标点与符号的处理

泰文：

- 借鉴了傣文机器翻译系统的词法分析、句法分析和语料预处理方法[3]。

汉语：

首先，对汉语语料中的非汉字字符进行统一化处理；然后，使用分词工具对汉语语料分词。

- 行尾转换
- 全角转半角
- 转义字符的处理
- 汉语分词
- 领域分类
- 过滤过长的句子或者长度不匹配（相差过大）的句子
- Tokenisation 处理
- 标点与符号的处理

5.2 分词与分句

分词方面，在汉语上，我们使用了基于汉语词库的分词分字工具；在蒙语上，我们使用了基于蒙语词库的短语分词以及词干分词工具；在藏语上，我们采用了西北民族大学李亚超的藏文分词技术[4]；在泰文上，我们使用了基于泰文词库的短语分词以及词干分词工具。

5.3 领域分类

本次评测语料均为综合领域。在模型预测时，我们做了人工判别工作,并重设调优参数阈值，在最终的文本分类结果文件中随机抽样并进行人工判断，预测结果的准确率在 90%以上。

5.4 词法分析技术

蒙古语为形态丰富的黏着语，在有限语料神经网络机器翻译系统中数据稀疏严重。我们对蒙古语进行联合词性标注的词法分析[5]，并采用基于规则的方法对切分结果进行后处理。我们了构建词级、词干级、词干词缀级神经网络机器翻译系统。

5.5 语言模型

对于泰中新闻,以该项目双语训练语料的目标语言作为一个集合使用了分类工具进行新闻/非新闻分类，新闻类语料训练得到的模型。对于藏汉机器翻译项目，使用该项目提供的藏汉语料作为训练集得到训练语言模型。对于蒙汉机器翻译项目，使用该项目提供的蒙汉语料作为训练集得到训练语言模型。

5.6 翻译模型训练

本次评测主要用到了基于短语的翻译模型和字词混合模型。模型主要包括以下几个模块：

1）预处理模块

根据各个语言的语言特点进行平行语料的预处理。

2）分词模块

采用最大似然方法，进行分词处理。

3）训练模块

搭建 TensorFlow 环境，语言模型使用 NMT 进行训练。

5.7 语料的使用

在训练集上，我们语料的使用情况，在 4.1 节已经做了详细介绍。对于受限语料，从评测组织方发布的泰中新闻语料句对中，对语料进行了相应修改，比如删除过长、过短内容等；在评测组织方发布的藏汉政府文献语料句对中，对语料进行了相应修改，比如删除过长、过短内容等；在评测组织方发布的蒙文日常用语语料句对中，删除过长、过短语料，保留并修改了语料。

5.8 调参与解码

提交的主系统在 MC、TC 和 ThaiC 评测中均使用以下特征：

1. 前向语言模型
2. 后向语言模型
3. 翻译概率
4. 词汇化概率
5. 反向翻译概率
6. 反向词汇化翻译概率
7. 规则惩罚
8. glue 规则惩罚
9. 词惩罚

NMT 训练过程中修改参数阈值限制，获取最佳效果。

6 评测

6.1 评测环境

1） 硬件环境： Intel(r) Core(TM) i7-4790 3.60 GHz, 4 cores. Mem: 28G HDD: 1T

2） 软件环境： Operating System: Linux CentOS 7

6.2 蒙汉机器翻译评测

在蒙汉机器翻译评测中，开发集采用评测组织方提供的语料进行调参，我们在 primary 主系统上进行了实验。主系统采用训练 6 轮的 NMT 模型。

6.3 藏汉机器翻译评测

在藏汉机器翻译评测中，开发集采用评测组织方提供的语料进行调参，我们的所有系统均采用所有的 11 个特征，区别在于使用的参数有所不同。我们在 primary 主系统上进行了实验。主系统采用训练 4 轮的 NMT 模型。

6.4 泰中机器翻译评测

在泰中机器翻译评测中，开发集采用评测组织方提供的语料进行调参，区别在于使用的参数有所不同。我们在 primary 主系统上进行了实验。主系统采用训练 7 轮的 NMT 模型。

泰中机器翻译非受限评测结果比较理想，在主系统评测结果中，总共 8 家参评单位,提交了 11 个评测结果，本单位得分第 1，BLEU5-SBF=0.3546，BLEU5=0.3662。

6.5 结论

潍坊北大青鸟华光照排有限公司在泰中机器翻译非受限评测中效果较好，主要得益于 20 多年来我们在傣文信息化和傣文机器翻译方面所做的大量前期工作，但还存在一些不足。

分析以及整个参评过程，我们有以下经验：

1）系统方面：机器翻译系统还没得到更好优化。比如分词系统及各个子模型的参数，都可以得到优化，以此达到更高翻译效果，并能取得更好的成绩。

2）数据的预处理方面：我们后来发现训练数据中有很多错误。在提交最终翻译结果后，我们利用自己研发的基于该文种的拼写形式语言的拼写检查软件对该文种训练语料进行了拼写检查预处理，纠正了拼写错误。使用预处理过的训练数据之后，翻译效果有所提高。

3）速度方面：我们对两种神经网络架构，一个是 RNN（循环神经网络），另一个则是 CNN（卷积神经网络）进行的比对，其结果是 RNN 神经网络架构针对小规模语料翻译结果要优于 CNN，但在翻译速度方面 CNN 占优势。

RNN 机器翻译按照序列进行工作，也就是和人一样，按照顺序一个个的进行翻译。但要记住的一点是，目前比较主流的 GPU 最大的优点是可以进行并行计算。这样一来 RNN 就没法最大化利用 GPU 的计算能力。

而 CNN 则可以同时处理多个语言片段，并且具有信息分层处理能力。将文本序列化、单词向量化，经过分层处理后再输出结果。在分层过程中，还会不断回顾源文本来确定下一个输出序列，所以 CNN 在翻译解码速度方面占优。

6.6 评测分析

在本次评测中，组织方提供了测试语料。我们在 primary 主系统上进行了实验。结果表明，在语料层面，因为 primary 系统使用了较多的训练语料，涉及领域相对非受限系统来讲范围相对缩小，因此翻译质量比想象中的好；在模型层面，由于我们的机器翻译系统集成了自主研发的语料处理模块，primary 给出的翻译结果均优于传统的翻译模型。对译文进行分析，由于我们的翻译系统都是基于词干，因此，与基于词的传统模型相比，翻译结果中未登录词所占比例较小。

7 致谢

本项目借鉴了厦门大学苏劲松“基于双语对应递归自编码器的汉藏统计机器翻译模型研究与实现”（CCF 开放课题，本公司资助）和西北民族大学李亚超“藏汉在线互译系统研究与实现”（CCF 开放课题，本公司资助）的部分成果。

本项目得到呼和浩特民族学院斯日古楞教授、包乌格德勒副教授和西双版纳报社玉康龙副总编等多位专家的指导。

在此一并表示感谢！

参考文献

[1]金卓林. 结合语义向量的双向机器翻译模型及评价[D].哈尔滨工业大学,2020.
[2]牛向华. 基于单语语料库训练的蒙汉机器翻译的研究[D].内蒙古工业大学,2019.
[3]殷建民.傣文信息技术研究进展[J].广西科学院学报，2018.34（1）：12-17,26.
[4]李亚超,江静,加羊吉,于洪志.TIP-LAS:一个开源的藏文分词词性标注系统[J].中文信息学报,2015,29(06):203-207.
[5]玉霞.蒙古文词法分析及其在蒙汉统计机器翻译中的应用[D].内蒙古师范大学，2015.