深度学习在汉藏机器翻译中的应用研究

李博涵[1,2]，刘汇丹[1]，龙从军[1,3]，吴健[1]

（1. 中国科学院软件研究所，北京 100190；2. 中国科学院大学，北京 100049；
3. 中国社会科学院民族学与人类学研究所，北京 100081）

摘　要：该文将深度学习技术应用于汉藏机器翻译任务中，采用了编码器-解码器结构。在编码阶段，首先将汉语句子中的每个词映射为定长的词向量，并通过循环神经网络压缩整个句子的全部信息。在解码过程中引入注意机制，使得解码器更集中的去注意当前翻译词的上下文依赖词，并每次选择概率最大的翻译词生成目标句子。该文使用上述方法在以法律文本、政府公文、新闻为主的书面语语料和口语类语料上进行实验。实验数据表明，在书面语语料上，NIST 和 BLEU 值分别达到了 6.39 和 0.296，在口语语料上，NIST 和 BLEU 值分别达到了 5.41 和 0.222。在相同的语料上，以短语为基本单元的 Moses 翻译模型的性能，书面语语料 NIST 和 BLEU 值为 6.98 和 0.335，口语语料 NIST 和 BLEU 值为 6.24 和 0.272。

关 键 词：深度学习；机器翻译；编码器-解码器；注意机制；

中图分类号：TP391　　**文献标识码：**A

Research on the Application of Deep Learning in Chinese-Tibetan Machine Translation

LI Bohan[1,2], LIU Huidan[1], LONG Congjun[1,3], WU Jian[1]

（1. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081, China）

Abstract: In This paper, we built a Chinese-Tibetan machine translation system based on encoder-decoder structure. The encoder encodes a source sentence into a fixed-length vector by using recurrent neural network. The decoder generates a translation word by word and allows a model to automatically search for parts of a source sentence that are relevant to predicting a target word by using the mechanism of attention. This approach is used to experiments on written style corpus, including legal text, government documents, news, and spoken style corpus. Experiments show that the method achieves a NIST score and BLEU score of 6.39 and 0.296 on a written style corpus, and 5.41 and 0.222 on a spoken style corpus. While on the same corpus, based on Moses phrase model, it achieves a NIST score and BLEU score of 6.98 and 0.335 on a written style corpus, and 6.24 and 0.272 on a spoken style corpus.

Key words: deep learning; machine translation; encoder-decoder; attention;

1. 引言

深度学习是一种基于深层神经网络(Deep Neural Network)的学习方法。从结构上说，深层神经网络就是前向多层神经网络，是由沃伦·麦卡洛克在 1943 年提出[1]。近年来，深度学习技术在自然语言处理领域也越来越受到重视，并逐渐应用于分词、词性标注等自然语言处理的任务中。而在机器翻译任务中，深度学习技术也已经成功的运用到了英法、汉英机器翻译中。本文主要研

基金项目：国家自然科学基金资助项目（61303165，61540057，61132009）；青海省自然科学基金资助项目（2016-ZJ-Y04，2016-ZJ-740）

作者简介：李博涵（1991--），男，硕士生，主要研究方向为多语言处理；刘汇丹（1982--），男，副研究员，主要研究方向为操作系统与多语言处理；龙从军（1978--），男，副研究员，主要研究方向为藏语计算语言学；吴健（1962--），男，研究员，主要研究方向为操作系统与中文信息处理。

究将深度学习技术应用于汉藏机器翻译。

本文接下来将从以下几个方面进行介绍，第 2 节主要对汉藏机器翻译及深度学习应用于机器翻译的研究现状进行介绍。第 3 节我们对介绍深度学习应用于汉藏机器翻译的方法。第 4 节给出实验结果及相关分析。文章最后对本文的相关工作进行总结。

2. 研究现状

机器翻译的研究主要包括基于规则和基于统计的机器翻译等。目前研究的热点主要是基于统计的机器翻译方法上。一般来说，基于统计的机器翻译需要构建较大的平行语料库，而少数民族语言资源相对匮乏，所以少数民族语言的统计机器翻译还有很多问题亟待解决。

2.1 汉藏机器翻译的研究现状

目前，汉藏机器翻译方法可以分为基于规则的方法和基于语料库统计的方法。

针对基于规则的方法，德盖才郎等提出了一种基于规则知识库的汉藏机器翻译系统[2]，才藏太构建了班智达汉藏公文规则翻译系统，提出将词条和语法相结合的翻译方法[3]。扎洛等提出了针对汉藏翻译中复句的翻译规则[4]。看卓才旦等从藏文的特性出发，研究了汉藏机器翻译中动词处理的方法[5]。张国喜针对汉藏机器翻译中的人名翻译问题进行了探讨[6]。这些基于规则的方法利用了藏文的本身特性，具有翻译速度快，实用性高的特点，但是构建和维护翻译规则库需要大量的人工资源和语言知识，工作量太大。

针对基于统计语料库方向上的研究，诺明花等尝试藏文词串频率统计算法和序列相交算法，有效抽取出平行语料中的汉藏短语对[7]。熊维等提出了一种基于短语串实例的翻译方法，利用词语对齐方法挖掘现有平行语料的资源，检索出任意长度短语串的翻译实例[8]。Huidan Liu 等提出了基于条件随机场的藏文分词方法，有效提高了藏文分词的效果[9]。Xin Yu 等提出基于词典的汉藏双语句子对齐方法[10]。这些分词、短语抽取、句子对齐等技术的研究都是基于语料库的汉藏统计机器翻译研究的基础。，基于统计的方法不需要耗费巨大的人力维护规则库，并且也取得了不错的效果。

2.2 深度学习应用于机器翻译的研究现状

统计机器翻译是深度学习近几年来应用于自然语言处理任务的热点问题之一。

Kyunghyun Cho 等提出了一种基于编码器-解码器的方法[11]来解决机器翻译问题。其中，编码器将输入序列通过循环神经网络[12]压缩成一个定长的向量，这个向量包含输入序列的全部信息。解码器将这个向量解码，生成与源语言句子对应的目标语言句子。编码器-解码器的方法在一定程度上能够解决机器翻译问题，但是当处理的训练语料规模很小的时候，会出现翻译质量随着源语言句子长度增加而急剧下降的问题。

针对这个问题，Bahdanau 提出了软注意机制[13]，软注意机制不再将源语言压缩成一个定长的向量，而是依据上下文依赖词来编码向量，借助这种变长的表示方法，解码器能够针对每个目标词选择性的集中在一个或者多个上下文依赖的词表示上。

3. 翻译方法

我们采用深度学习技术，构建了一个汉藏机器翻译系统。系统的处理流程主要包括以下几个步骤：一是针对汉藏双语句子对平行语料，分别采用斯坦福分词器[1]和卓玛拉藏文词法分析工具[2]对汉语、藏语句子分词。二是将分词后的句子序列输入编码器-解码器模型，并引入注意机制，进行模型的训练。三是将汉语测试语料输入到编码器-解码器模型中，从而得到最终的翻译结果。

3.1 系统流程

[1] http://nlp.stanford.edu/software/segmenter.shtml

[2] http://www.tibetannlp.cn/

整个系统的处理流程如图 1 所示。

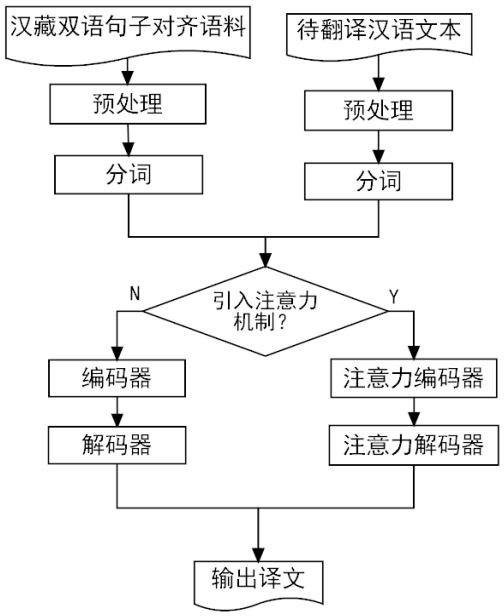

图 1　汉藏机器翻译系统流程图

3.2 简单编码器-解码器结构

1）编码器：

首先我们将源语言句子中的每个词语分别映射为一个 N 维词向量 w_t，将每一个词向量通过矩阵 E 映射为一个隐层状态 h_t。然后通过循环神经网络，将源语言句子中的所有信息压缩成一个向量 h_T ，这个向量包含整个源语言句子的全部信息。

由于传统的循环神经网络会出现梯度爆炸或梯度消失的问题[14]，这里我们采用 GRU 单元[11]作为神经网络的节点单元。

这个过程可以表示为图 2：

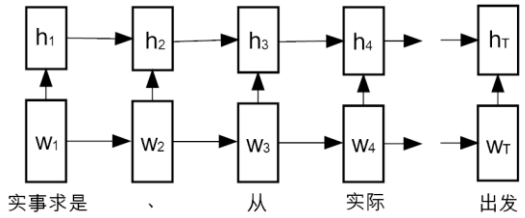

图 2 编码器过程表示

图中隐层状态 h_t 的计算方法如式子(1)：

$$h_t = \mathrm{f}(h_{t-1}, w_t) \qquad (1)$$

其中 f 是激活函数，通常采用 Sigmoid 函数，h_0 是零向量，那么最终计算出来的隐层状态 h_T 便可以用来表示整个句子的全部信息。

2）解码器：

通过上述的编码器，我们可以得到一个包含整个源语言句子信息的向量。而解码器的过程表示如图 3 所示：

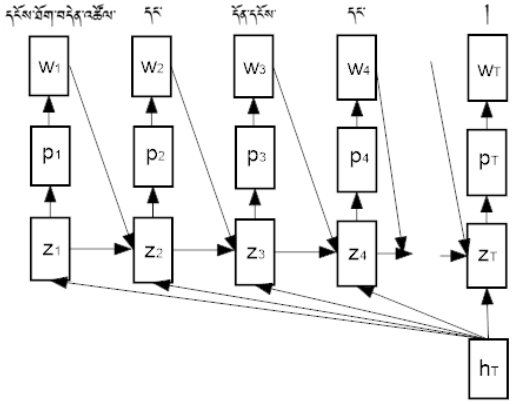

图 3 解码器过程表示

图中 h_T 表示编码器得到的向量，z_t 表示解码器方向的目标语言句子每个词的隐层状态，可以通过前一个隐层状态 z_{t-1}，前一个词向量 w_{t-1} 和 h_T 计算出来，如式子(2)：

$$z_t = \mathrm{g}(h_T, z_{t-1}, w_{t-1}) \qquad (2)$$

其中 g 表示激活函数，通常为 Sigmoid 函数。

通过目标句子每个词的状态，可以计算出目标句子每个词的概率分布 p_t，那么通过 Softmax 函数可以计算出每个词所取概率的最大那个词，即为目标预测词。计算方式如式子(3)：

$$\mathrm{p}(y_t | y_{<t}, \mathbf{x}) = \mathrm{Softmax}(\mathrm{W}z_t + b_z) \qquad (3)$$

其中 W 维度为$|V| \times H_z$，$|V|$表示目标语言词典的大小，H_z 为隐层状态的维度。Wz_t+b_z 就是对每个可能的输出单词打分，通过 Softmax 归一化可以得到 y_t 为目标词典中每个词的概率大小，选取概率最大的那个词作为译词。

通过式子(3)的计算方式生成目标句子后，再通过极大似然估计的方法训练模型。在训练过程中，我们最大化的目标是给定一个源语言句子 x，输出正确目标句子 y 的概率：

$$L(D,\theta)=\frac{1}{N}\sum_{n=1}^{N}\log\left(p(y^n|x^n,\theta)\right)=$$

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{Tn}\log\left(p(y_t{}^n|y_{<t}^n,x^n,\theta)\right) \quad (4)$$

针对似然函数 L 采用随机梯度下降法（Stochastic Gradient Descent），不断迭代寻找最优参数，直到训练过程收敛。

3.3 注意机制

简单编码器-解码器可以解决机器翻译的问题，但也有一定的不足。在训练过程中，简单编码器会将任意长度的源语言句子压缩成定长的向量，那么更长的句子就会面临信息损失的问题。

基于此，我们采用了注意机制来更好地解决这一问题。每当解码器对每个目标词进行解码时，解码器会更集中的去注意这个词的上下文依赖词，而不是整个句子，由于传统的单向循环神经网络只考虑了前面状态 $h_{<t}$ 对当前状态 h_t 的影响，故这里我们将之前的编码器改用双向循环神经网络[15]来将源语言句子编码压缩，双向循环神经网络不仅可以考虑前面状态 $h_{<t}$ 对当前状态 h_t 的影响，也考虑了后面状态 $h_{>t}$ 对当前状态 h_t 的影响情况。注意机制的过程如图 4 所示。

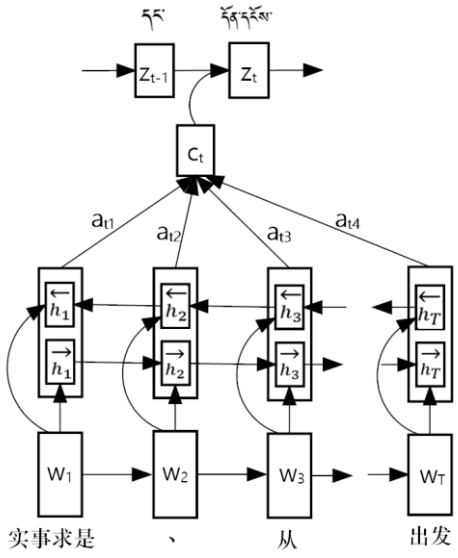

图 4 注意机制过程表示

其中，上下文状态 c_t 表示第 t 个词的上下文信息。计算方法如式(5)所示：

$$c_t=\sum_{t=1}^{T}a_{tj}h_j \quad (5)$$

式子(5)中的 a_{tj} 的计算方法，见式子(6)：

$$a_{tj}=\frac{\exp(e_{tj})}{\sum_{t=1}^{T}\exp(e_{tk})} \quad (6)$$

式子(6)中 e_{tj} 的计算是基于解码器中的前一个隐层状态 z_{t-1} 和源语言上下文依赖 h_j 作为输入，如式子(7)：

$$e_{tj}=q(z_{t-1},h_j) \quad (7)$$

4. 实验方法与结果分析

下面，我们通过实验来考察引入注意机制的编码器-解码器的翻译效果，为了验证这个方法的通用性和有效性，我们分别采用了书面语语料和口语类语料来进行验证，并与开源机器翻译系统 Moses[3]翻译出来的结果进行对比。

实验所用的语料情况如下：

语料 1：中国科学院软件研究所汉藏双语平行语料，包括政府新闻类、法律类、伟人著作等，共 39 万句对训练语料，1000 句测试语料。

语料 2：中国社会科学院民族学与人类学研究所藏汉双语口语类平行语料，其中训练语料 15 万句对，测试语料 1000 句。

最终的翻译质量采用基于词的 NIST[16]和 BLEU[17]值来评估。

4.1 实验方法

编码器-解码器方法：针对训练语料中的汉语句子采用斯坦福分词器分词，藏语句子采用卓玛拉藏文词法分析工具分词。然后将分词后的训练语料输入到引入注意机制的编码器-解码器模型中进行训练。最后将待翻译测试语料输入系统，生成最终的翻译译文。

基于短语模型的翻译方法：采用 Moses

[3] http://www.statmt.org/moses

系统中的基于短语的翻译模型，在同样的语料进行训练和测试，得到翻译译文。

4.2 性能与分析

我们采用上述两种方法，分别在语料 1 和语料 2 上进行实验，统计得到的 NIST 值和 BLEU 值如表 1。

表 1 NIST 和 BLEU 统计结果

指标	NIST		BLEU	
语料	语料 1	语料 2	语料 1	语料 2
En-De	6.39	5.40	0.296	0.222
Moses	6.98(0.59↑)	6.24(0.84↑)	0.335(0.039↑)	0.272(0.05↑)

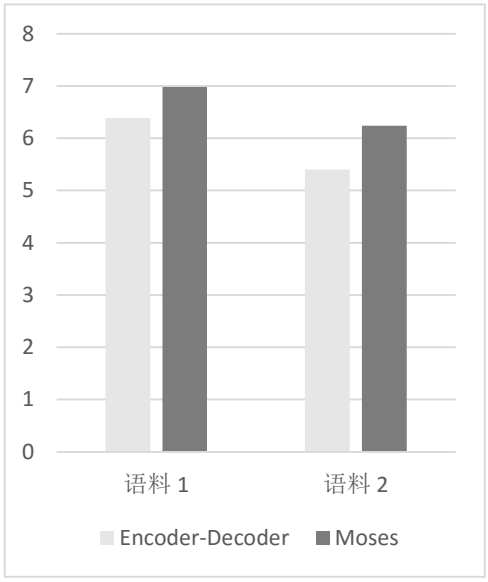

图 5 NIST 统计结果

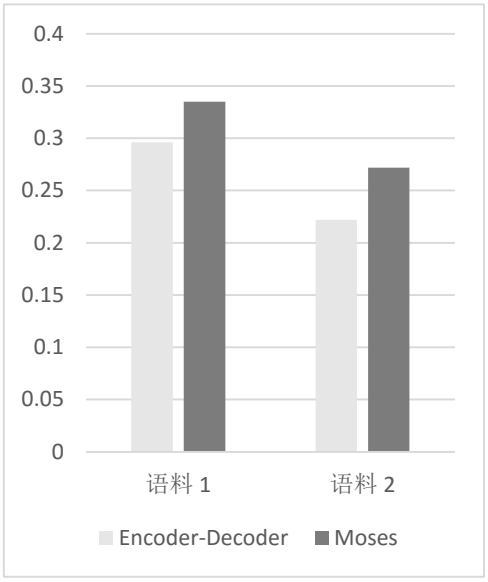

图 6 BLEU 统计结果

数据表明，这种基于编码器-解码器的深度学习方法，与传统的基于短语模型的方法在翻译质量上还有一些差距。经过分析，原因可能有以下几点：

一是语序调整的问题：

例如针对汉语句子“我常讲，我们大家都在建设有中国特色社会主义事业的航船上。”。

标准答案为：

“ངས་རྟག་ཏུ་ཚོ་ཚང་གྱང་གིའི་ཞུན་ལེས་ལྟན་པའི་སྒྲིག་ཚོགས་རིང་ལུགས་ཀྱི་བྱ་གཞན་གཞོགས་སྐྱེན་བྱེད་པའི་ཐུ་གཅེན་སེ་ཡོད་ཅེས་བཤད་ཀྱིན་ཡོད།”。

编码器-解码器方法给出的答案是：

“ངས་（我-施格）རྟག་ཏུ（经常）གྱང་གིའི་（中国的）ཞུན་ལེས་（特色）ལྟན་པ（具有）འི（的）སྒྲིག་ཚོགས་རིང་ལུགས་（社会主义）ཀྱི（的）བྱ་གཞན་（事业）འཛུགས་སྐྱེན（建设）ཞེད་པ（做）འི（的）ཐུ་གཅེན་（航船） སྟེང་དུ（上在）ཚང་མས་（大家-施格）”。

Moses 给出的答案为：

“ངས་（我施格）རྒྱུན་དུ（经常）བཤད་ཀ（说的话）ང་ཚོ་（我们）ཚང་མ་（大家）གྱང་གིའི་ཞུན་ལེས་ལྟན་པའི་སྒྲིག་ཚོགས་རིང་ལུགས་（中国特色的社会主义）ཀྱི་བྱ་གཞན་（事业）འཛུགས་སྐྱེན་ཞེད་པའི་（建设的）ཐུ་གཅེན་སྟེང་།（航船上）”。

值得注意的是，编码器-解码器方法给出译文中的最后一个词“ཚང་མས་（大家）”按照语序应该置于句子中部，而基于短语的模型则得到了正确的语序。

二是语法关系清晰度：

针对汉语句子“（二）低于当地最低工资标准支付劳动者工资的”。

标准答案为：

“ (གཉིས།) ངལ་རྩོལ་པར་ས་གནས་དེ་གའི་བླ་ལོགས་དམའ་ཤོས་ཀྱི་ཚད་གཞི་བས་དམའ་བའི་བླ་ཆས་སྦྱོད་རིགས།”

编码器-解码器方法给出的答案是：

“（གཉིས）（二）ས་གནས་དེ་གའི་（这些地方的）གླ་ཡོག་དམའ་ཤོས་ཀྱི་（最低工资）ཚད་གཞི་（标准）ལས་（比）དམའ་བ་（低）དང་（和）རལ་སྐྱེས་པའི་（劳动者的）གླ་ཆ（工资）སྦྱོར་རྒྱུ་བཅས་ཡིན།（给等是）”。

Moses 给出的答案是：

“（གཉིས）（二）ས་གནས་དེ་གའི་（这些地方的）གླ་ཡོག་དམའ་ཤོས་ཀྱི་（最低工资）ཚད་གཞི་（标准）ལས་（比）དམའ་བའི་（低的）རལ་སྐྱེས་པའི་གླ་ཆ（劳动者的工资）སྦྱོར་དགོས།（要给）”。

这个句子中“རལ་སྐྱེས་པར་（劳动者的）”是对象宾语，直接宾语的中心名词是“གླ་ཆ（工资）”，受到带格标记的短语“ས་གནས་དེ་གའི་གླ་ཡོག་དམའ་ཤོས་ཀྱི་ཚད་གཞི་ལས་དམའ་བའི་（低于当地最低工资标准的）”的修饰，语法关系比较清晰。如果从语法关系这个角度来分析两个系统的翻译结果，发现 Moses 系统更加接近标准答案，满足标准答案的修饰关系。而编码器-解码器系统翻译结果中，由藏文标点符号“།”分成两个短语，但是两个短语之间关系并不明确，无法体现出标准答案中的中心名词“གླ་ཆ（工资）”受到带格标记短语“ས་གནས་དེ་གའི་གླ་ཡོག་དམའ་ཤོས་ཀྱི་ཚད་གཞི་ལས་དམའ་བའི་（低于当地最低工资标准的）”修饰的关系。

我们认为产生这些现象的根本原因，主要在于编码器-解码器的方法还是基于词的翻译，而传统的方法则采用了基于短语的翻译。编码器-解码器的这种翻译方式使得待翻译的每个句子序列中单元数量更多，故单元之间的组合方式更多，翻译结构相对复杂，词语之间的语序更加难以调整。而基于短语的翻译方法好处在于短语模型只关注于短语间的语序调整，并且保证每个短语内部是符合语法关系的，故在一定程度上避免了上述问题。

5. 结束语

本文将深度学习技术应用于汉藏机器翻译任务中，并采用了编码器-解码器结构。编码器首先将汉语句子中的每个词映射为定长的词向量，通过循环神经网络压缩整个句子的全部信息。解码器中引入了注意机制，即根据当前词的上下文依赖词，每次选择概率最大的译词，生成目标句子。实验数据表明，在书面语语料上，NIST 和 BLEU 值分别达到了 6.39 和 0.296，在口语语料中，NIST 和 BLEU 值分别达到了 5.41 和 0.222。而以短语为基本单元的 Moses 翻译模型上，书面语语料 NIST 和 BLEU 值为 6.98 和 0.335，口语语料 NIST 和 BLEU 值为 6.24 和 0.272。

参考文献

[1]McCulloch, Warren S and Walter Pitts. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics.1943:115-133.

[2]德盖才郎,李延福,项青朝加等.实用化汉藏机器翻译系统的设计与实现[C]//863 计划智能计算机主题学术会议论文集.2001:405-411.

[3]才藏太,华关加.班智达汉藏公文翻译系统中基于二分法的句法分析方法研究[J].中文信息学报.2005:7-12.

[4]扎洛,索南仁欠.汉藏机器翻译中复句的翻译规则研究[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议.2006:454-460.

[5]看卓才旦,金为勋,李延福等.汉藏翻译系统中的动词处理研究[J].术语标准化与信息技术.2006:28-32.

[6]张国喜.汉藏机器翻译中人名翻译问题探讨[C]//首届少数民族青年自然语言信息处理学术研讨会.2004:12-14.

[7]诺明花,吴健,刘汇丹等.汉藏短语对抽取中短语译文获取方法研究[J].中文信息学报.2011:112-117.

[8]熊维,吴健,刘汇丹,张立强.基于短语串实例的汉藏辅助翻译[J].中文信息学报.2013:84-90.

[9]Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu and Yeping He.Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Fields[C]//In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation(PACLIC-2011):168-177.

[10]Xin Yu, Weina Zhao, Jian Wu. Dictionary-based Chinese-Tibetan sentence alignment[C]//The 2010 IEEE International Conference on Intelligent Computing and Integrated

Systems.2010:489-493.
[11]Kyuunghyn Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio.On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J]. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.2014:103-111.
[12]Tsoi A C. Recurrent neural network architectures: An overview[J].Lecture Notes in Computer Science.1998:1-26.
[13]Dzmitry Bahdanau, Kyung Hyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate[J].Computer Science.2014.
[14]Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J].IEEE Transaction on Neural Networks.1994:157-66.
[15]Schuster M. and Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transaction on Signal Processing.1997: 2673－2681.
[16]Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics[C]//International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. 2002:138-145.
[17]Kishore Papineni, Salim Roukos, Todd Ward, WeiJing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation[J]. Proceedings of Annual Meeting of the Acl.2002: 311-318.