基于子字单元的神经机器翻译未登录词翻译分析[1]

韩冬，李军辉，熊德意，周国栋

（苏州大学计算机科学与技术学院，江苏省苏州市 215006）

摘　要：神经机器翻译(NMT)为机器翻译系统提供了一种全新的方法，它与传统的统计机器翻译系统（SMT）相比，翻译结果具有更加流畅的优势。但是 NMT 系统也有着其自身的缺点：翻译精准度的问题，尤其是对未登录词的翻译。2016 年，Rico Sennrich 和 Barry Haddow 等人提出了 Byte Pair Encoding (BPE)的方法，将原有的单词拆解成了更小单元的高频词进行翻译。现如今，这种方法已经被广泛用于各种开源的神经机器翻译系统中。本文主要针对 BPE 方法在中英神经机器翻译中的应用，分析 BPE 方法在多大程度上解决了未登录词翻译的问题。实验表明，与传统的 NMT 系统相比，BPE 方法获得了 1.02 BLEU 值的提升，对未登录词的翻译精准度达到了 45%左右，与 SMT 系统翻译精准度相似。因此可以得出结论：BPE 的方法是一种对 NMT 系统中未登录词问题的行之有效的解决方法。

关键词：神经机器翻译；BPE 编码；未登录词

中图分类号：TP391 **文献标识码：**A

An Experimental Analysis of Unknown Words in Neural Machine Translation Using Sub - word Unit

Dong Han, Junhui Li, Deyi Xiong, Guodong Zhou

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu, 215006, China)

Abstract: Neural Machine Translation (NMT) provides a new way for machine translation systems to have a smoother advantage over traditional statistical machine translation systems (SMT). But NMT system also has its own problems: the translation of unknown words. Recently Sennrich and Haddow (2016) proposed Byte Pair Encoding (BPE) method for NMT, which disassembles the original word into smaller units of high frequency sub-units. Afterwards, this BPE method has been widely applied in various open source NMT systems. This paper tries to answer the question that how well does BPE method resolve the translation of unknown words in Chinese to English translation. Experiments show that compared BPE method achieves 1.02 BLEU improvements over conventional NMT system. Further analysis reveals that BPE method has correctly translated 45% unknown words, which is similar to that of the SMT system. Based on our experiment analysis, we conclude that BPE method is an effective solution to the problem of unknown words in NMT systems.

Keywords:Neural Machine Translation ; BPEencoding; Unknownwords

1 引言

神经机器翻译（NMT）是一种新颖的方法来解决机器翻译问题，并且最近几年已经取得了极大的成功[1-3]。尤其是在翻译的流利度方面，NMT 系统与传统的 SMT 系统相比，翻

[1]**收稿日期： 定稿日期：**

基金项目：国家自然科学基金（61401295）

作者简介：韩冬（1993——），男，硕士研究生，主要研究方向为自然语言处理，机器翻译；李军辉（1983——），男，副教授，主要研究方向为自然语言处理，机器翻译，句法分析；熊德意（1979——），男，教授，主要研究方向为机器翻译，自然语言处理；周国栋（1967——），男，教授，主要研究方向为自然语言处理。

译结果更加顺畅。总的来说，NMT 系统采用神经网络的结构，其不需要去存储短语表，而是有着一个小规模的词汇表，这大大减小了计算的复杂度。

但是，NMT 系统也有着自身的缺点。因为 NMT 系统为了能够控制计算的复杂度，有着一个固定大小的词汇表，通常会将词汇表限制在 30k 到 80k 之间，这就导致了其在翻译未登录词时有着严重的不足。由于限定词汇表的大小，对于未出现在该词汇表中的词，NMT 系统用 UNK 标记来替代。结果，NMT 系统不仅无法将它们翻译准确，而且破坏了句子的结构特征。为了解决 NMT 系统中存在的这一问题， Sennrich 和 Haddow(2016)【4】提出了一种 BPE 编码[5]的解决方法。该方法将训练语料中单词拆分成更为常见的小部分，这里把它叫做子字单元。通过这种方法，我们假设在同样将词汇表设置成 30k 的情况下，由于很多单词拆解的子字部分是相同的，所以 30k 的子字单元实际上可以表示出远远超出 30k 的以单词为基础的词汇表。这样，对于绝大多数未登录词，就可以通过子字单元的组合表示出翻译的结果。

将单词拆解为子字单元的方法对于未登录词问题确实是一种简单的方法，但是对于其翻译的效果我们依然持疑问的态度。因此，本文对 BPE 方法的翻译结果进行了分析。分析 BPE 方法是如何翻译未登录词的，在多大程度上解决了 NMT 系统对未登录词的翻译问题。

通过对中英文翻译的实验结果分析，本文有如下发现：

- 验证了 BPE 方法对未登录词确实是一种行之有效的方法，在对中英文双向都拆解成子字单元的实验中，实验结果与不做处理的 NMT 系统相比，提高了 1.02 BLEU 值。
- 本文进行了四组实验，对各个实验中训练语料中未登录词进行了统计，发现通过 BPE 方法的实验在训练语料中基本涵盖了所有的训练单词。
- 统计了各测试语料中源端未登录词的个数，然后得出结论：使用中英均做 BPE 的方法，测试源端语料基本不会出现未登录词。从翻译结果看，目标端翻译结果中不含有 UNK 标识符。从而可以说明通过 BPE 的方法确实极大程度上解决了未登录词的问题。
- 分析测试源端中未登录词的词性和各组对比实验的解决效果，发现 NMT 系统中未登录词的来源主要是名词，动词和数词。
- 与 SMT 方法比较，BPE 方法对未登录词的翻译效果在精准度上基本上保持一致。对于测试源端语料中未登录词的翻译均达到了 45%左右的正确率。

2 神经机器翻译系统

本节将简要的介绍本文的神经机器翻译系统。

本文的编码-解码 NMT 系统是基于注意力机制的循环神经网络[6]，根据源端句子的输入计算翻译结果的条件概率。在编码时，使用具有 GRU 单元的双向循环神经网络【7】，其正向和反向读入输入的序列 $X=(x_1,x_2,.....,x_m)$ 并且输出正向隐藏状态序列 $(h_1,h_2.....h_m)$ 和反向隐藏状态序列 $(h_1,h_2.....h_m)$，然后将这两个序列融合成为一个新的序列 $([h_1,h_1],[h_2,h_2].....[h_m,h_m])$。

解码器使用基于注意力机制的循环神经网络去预测目标端序列 $Y=(y_1,y_2,.....y_n)$，每个单词 y_j 通过隐藏状态 s_j，预测的前一单词 y_{j-1} 和一个上下文向量 c_j 所决定。

3 BPE[2]编码与子字单元

虽然有大量的工作用来不断地优化神经机器翻译系统，但是对于未登录词的解决仍然是现如今神经机器翻译系统的一大难题。

BPE 的思想是将单词拆解为更小更常见的子字单元。对于原本不在词表中的单词，NMT 系统一般会用 UNK 标示符替代。BPE 方法将其拆解为常见的子字，通过翻译子字部分将原

[2]可以在 http://aclweb.org/anthology/attachments/P/P16/P16-1162.Software.zip 得到 BPE 方法使用的代码

有的 UNK 单词进行了翻译，从而极大地保存了句子的结构特征和流畅性。

BPE 方法拆解成子字单元的具体效果可以通过下面的例子来进行说明：

(a) he is a good boy

(b) h@@ e is a g@@ o@@ o@@ d b@@ oy

假设句子（a）出现在训练语料中，传统的 NMT 系统在形成词汇表时，使用的是以单词为基础的划分方式，然后取词频出现较高的单词形成字典。但是 BPE 方法则是以一种介乎单词和字母之间的子字单元形成字典，如（b）所示，其将一个句子中的单词拆分成了更小的部分 he ----->h@@ 和 e。原本以 he 形成字典的方式转变为 h@@ 和 e 两个字典。更加详细的说明可参见（Sennrich and Haddow, 2016)。

在中文语料中，假设“大学文凭”是一个中文未登录词，被标记为 UNK，通过 BPE 的方法，将“大学文凭”拆解为“大学”和“文凭”两个部分，而“大学”和“文凭”这两部分恰恰是在词汇表中，可以准确翻译，从而可以得到“大学文凭”的正确翻译结果，如下所示。

Eg. 大学文凭----->大学 文凭 ---------->university diploma

4 实验

本文针对中到英的翻译任务分析 BPE 方法对未登录词的翻译效果。为此，共准备了四组实验，每组实验的翻译性能采用评测标准 BLEU 值[8]。

训练集包含从 LDC 语料库中抽取的 1.25M 句对的中文到英文平行语料。选择 NIST MT 06 数据集作为开发集，NIST MT 02, 03, 04, 05, 08 作为测试集。

表 1 给出了本文进行的四组实验，其中 Baseline 系统将所有的中英文端未登录词都替代为 UNK 标记，BPE_cn、BPE_en、BPE_all 指分别在源端（即中文端）、目标端（即英文端）和两端（即中英文端）进行 BPE 子单元处理。

在实验中，设置隐层单元的个数为 1000，源端和目标端单词词向量（word_embedding）的维度为 620 维。神经网络用 Adadelta [9]模型更新参数。设置 batch_size 为 80。我们使用 GPU 去运行实验训练部分，提高实验运行的速度。

系统	中文词表	英文词表	描述
Baseline	30k	30k	Baseline 为原有的 NMT 系统，未进行任何处理
BPE_cn	30k	30k	只对中文端做 BPE 处理
BPE_en	30k	30k	只对英文端做 BPE 处理
BPE_all	30k	30k	中文端和英文端都进行 BPE 处理

表 1：四组实验及其描述.

4.1 训练语料中未登录词的统计

本文统计了上述四个实验中未登录词的个数，如表 2 所示，从表中可以看出：

- 在相同的训练语料下，通过 BPE 方法可以极大的减少未登录词的个数，系统中中文未登录词的个数从 174291(Baseline 系统)减少到 549(BPE_all 系统)，英文未登录词的个数从 75910(Baseline 系统)减少到 0(BPE_all 系统)。
- 在测试语料中，因为测试语料单词的个数要远远小于训练中出现单词个数，所

以我们有理由相信：在中英均做 BPE 处理的实验中，在测试时，源端将产生极少的未登录词，英文翻译结果中将会有极少的 UNK 标识符。这在我们之后的实验结果中也得到了验证。

系统	单词总个数	词表单词个数	未登录词
Baseline	中：204291 英：105910	中：30000 英：30000	中：174291 英：75910
BPE_cn	中：30566 英：104697	中：30000 英：30000	中：566 英：74697
BPE_en	中：201037 英：29294	中：30000 英：29294	中：171037 英：0
BPE_all	中：30549 英：29284	中：30000 英：29284	中：549 英：0

表 2：四组实验中在相同训练语料下，词形总个数，词表大小和训练语料中未登录词的统计结果。其中我们设置源/目标端句子最大长度为 50，超过 50 个单词的句子舍弃，BPE 在拆解成子字单元的过程中，会增加句子的长度[3]

4.2 测试集未登录词的统计

表 3 统计了各个测试集源端包含的中文未登录词的个数。

系统	NIST02		NIST03		NIST04		NIST05		NIST06		NIST08	
	S-W	S-U	S-W	S-U	S-W	S-U	S-W	S-U	S-W	S-U	S-W	S-U
Baseline	22639	975	24095	1283	49438	1808	29897	1410	38349	1862	32310	2195
BPE_cn	24416	6	26447	15	52750	17	32291	12	41843	12	35940	14
BPE_en	22639	975	24095	1283	49438	1808	29897	1410	38349	1862	32310	2195
BPE_all	24416	11	26447	15	52750	22	32291	12	41843	14	35940	17

表 3：S-W:测试集中源端（中文端）单词总个数，S-U:测试集中未登录词个数。注意：在对源端做 BPE 处理时，源端单词个数会增多。

表 3 的统计表明，在 Baseline 系统中，测试集源端存在 5%左右的未登录词。我们把这部分词称为 $V_{Baseline\_chn\_UNK}$。

[3]原 Baseline 在最大句子长度设置为 50 时，实际用于训练的语料行数为 1128660。
采用 BPE_all 方法也将最大句子长度设置为 50 的情况下，实际用于训练的语料行数为 1119600，两种情况下训练规模近似相同，仅仅减少了 0.8%。

通过上表的统计结果，可以发现经过 BPE 处理的实验与未经任何处理的 Baseline 实验相比，其测试集中含有很少的未登录词。特别地，对于 BPE_all 系统，由于源端和目标端同时分别采用了 BPE 编码，这样就使得在翻译时源端和目标端词汇表中的单词基本上完全覆盖了测试集中的单词，所以翻译结果中将基本不会含有未登录词[4]。在测试的时候，虽然 BPE_all 系统中源端含有很少的未登录词，但是 BPE_all 是否可以将 $V_{Baseline\_chn\_UNK}$ 翻译正确，翻译的质量如何，我们在下一节将进一步讨论。

4.3 Baseline 系统测试集源端 UNK 分析

本节分析了 BPE_all 实验对 $V_{Baseline\_chn\_UNK}$ 集合中单词的翻译效果。图 1 给出了该集合中单词按词性的分布统计，不难看出 Baseline 系统中源端未登录词主要是名词（约占 76%），然后是动词（约占 16%）和数词（约占 6%）。

表 4 统计了在 BPE_all 实验中，将 $V_{Baseline\_chn\_UNK}$ 集合中单词正确翻译的比率。根据参考数据集，以 NIST06 为例，人工分析了 NIST06 中 $V_{Baseline\_chn\_UNK}$ 共 1826 个词的翻译准确率。以名词为例，从表 4 可以看出，$V_{Baseline\_chn\_UNK}$ 总共包含名词 1455 个，其中 669 个翻译正确，占 46%。说明 BPE 方法对源端未登录词具有一定的翻译效果。

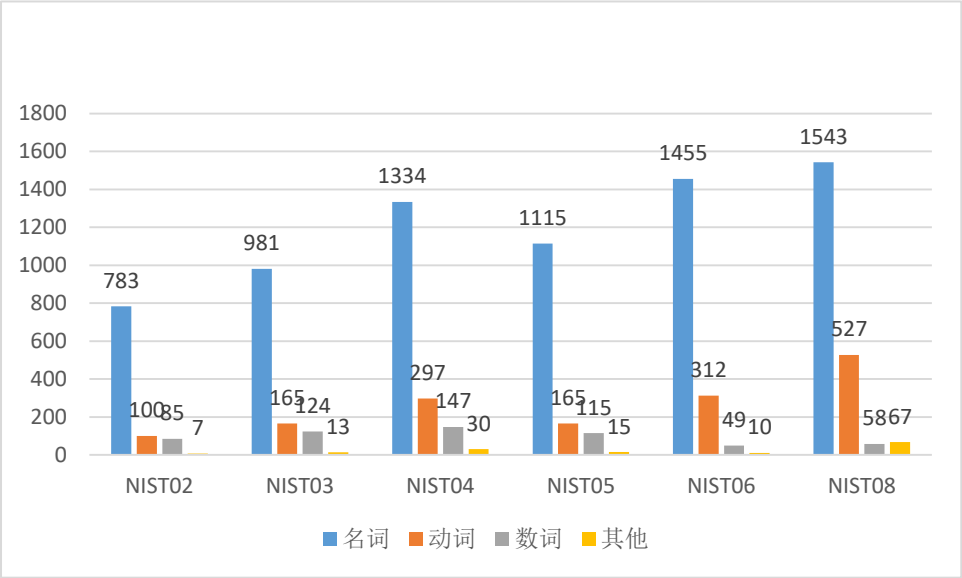

图 1：测试集 $V_{Baseline\_chn\_UNK}$ 的词性分布

名词		动词		数词		其他词性		所有词性	
总数	正确数	总数	正确数	总数	正确数	总数	正确数	总数	正确数
1455	669	312	125	49	32	10	4	1826	830（46%）

表 4：在 BPE_all 系统中，NIST06 开发集 $V_{Baseline\_chn\_UNK}$ 的翻译正确统计.

4.4 Baseline 系统译文 UNK 分析

本节分析在 Baseline 系统中，哪些源端词翻译为译文的 UNK。根据词对齐信息，找到译文中 UNK 所对应源端的单词，称源端的这些词为 $V_{Baseline\_to\_eng\_UNK}$。图 2 统计了 $V_{Baseline\_to\_eng\_UNK}$ 的词性分布情况，该图说明了导致译文中出现 UNK 最多的是名词(70%)，紧接着是动词(17%)和数词(8%)。表 5 统计了 $V_{Baseline\_to\_eng\_UNK}$ 中 UNK 是来自源端未登录

[4]如表 3 所示，在测试集中，源端仍存在极少数的未登录词，该未登录词基本上是特殊符号。

词的（$V_{Baseline\_chn\_UNK}$）个数，以及源端在词表中但最终翻译为 UNK 的单词个数。以 NIST02 为例，Baseline 的译文中共包含单词 25394 个，其中 UNK 的数量为 636 个。针对这 636 个译文 UNK，其中 534 个是由源端 UNK 翻译所至，另外 102 个是由源端非 UNK 翻译而来。这也说明了译文中产生 UNK 的单词大部分来自于源端的未登录词。

接着，由于目标端已消除 UNK（BPE_all 译文中没有 UNK 标示符），本节分析 BPE_all 系统又是如何翻译 $V_{Baseline\_to\_eng\_UNK}$ 中的词，其翻译质量又如何。以 NIST06 为例，本文人工分析了 NIST06 中 $V_{Baseline\_to\_eng\_UNK}$ 共 1204 个词的翻译准确率，如表 6 所示。以名词为例，从表 6 可以看出，$V_{Baseline\_to\_eng\_UNK}$ 总共包含名词 881 个，其中 396 个翻译正确，占 45%。

	NIST02	NIST03	NIST04	NIST05	NIST06	NIST08
译文单词数	25394	27697	55452	33339	42934	35293
UNK	636	887	1184	986	1204	1250
Chn_UNK	534	637	936	767	966	962
Non_Chn_UNK	102	250	248	219	238	288

表 5 ：测试集译文中 UNK 的统计。其中 UNK 行指译文中出现 UNK 的个数；Chn_UNK 行表示多少数量的译文 UNK 翻译自源端 UNK；Non_Chn_UNK 行表示多少数量的译文 UNK 翻译自源端非 UNK 词。

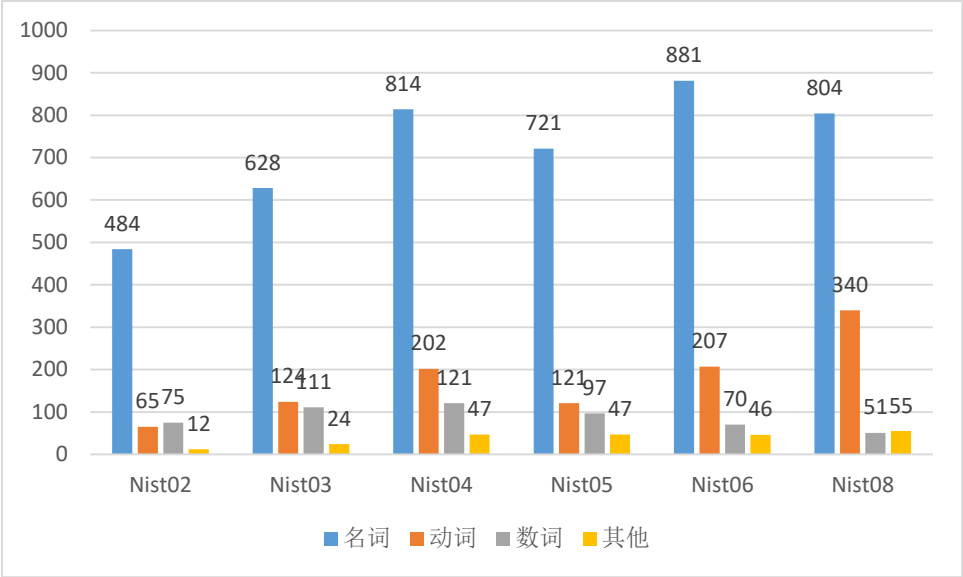

图 2：测试集 $V_{Baseline\_to\_eng\_UNK}$ 的词性分布

名词		动词		数词		其他词性		所有词性	
总数	正确数	总数	正确数	总数	正确数	总数	正确数	总数	正确数
881	396	207	98	70	46	46	7	1204	547（45%）

表 6 ：在 BPE_all 系统中，NIST 06 测试集 $V_{Baseline\_to\_eng\_UNK}$ 的翻译正确统计.

我们评测了各词性下的 UNK 单词在 BPE 翻译结果下的正确数，发现最终使用 BPE 方法的

实验可以将 45%的 UNK 单词正确翻译（其余单词翻译错误或漏翻），从而在一定程度上保障了句子的结构特征和流畅性。

4.5 BPE 与 SMT 比较

从以上分析可以看出，BPE 编码可以在一定程度上解决未登录词的翻译问题。但是，与 SMT 系统相比，BPE 是否能够更好的解决未登录词的翻译仍未知。由于 SMT 并没有限定词汇表，对 $V_{Baseline\_to\_eng\_UNK}$ 中的词的翻译效果要比 NMT Baseline 系统好。本节主要比较 SMT 系统[5]与 BPE_all 系统，分析两者对 $V_{Baseline\_to\_eng\_UNK}$ 中的词的翻译效果。

以 NIST06 为例，本文人工分析了 NIST06 中 $V_{Baseline\_to\_eng\_UNK}$ 共 1204 个词的在 SMT 系统下的翻译准确率，如表 7 所示。

名词		动词		数词		其他词性		所有词性	
总数	正确数	总数	正确数	总数	正确数	总数	正确数	总数	正确数
881	376	207	115	70	64	46	13	1204	568（47%）

表 7：在 SMT 系统中，NIST 06 测试集 $V_{Baseline\_to\_eng\_UNK}$ 的翻译正确统计.

比较表 6 和表 7，不难看出，BPE 方法和 SMT 系统在翻译精准度上基本持平，最终对 UNK 单词翻译的精准度均达到了 45%左右，从而可以说明 BPE 方法在一定程度上既具有传统 NMT 系统的流畅性，又具有接近 SMT 系统的未登录词翻译精准度。

4.6 主要结果

表 8 给出了 Baseline 系统和各 BPE 系统在测试集上的翻译性能 BLEU 值。从表 8 中可以看出，仅对源端或目标端采用 BPE 编码，能够在一定程度上提高翻译性，两端同时采用 BPE 编码，能进一步显著地提高翻译的性能，例如 BPE_all 系统在测试集上比 Baseline 系统平均提高了 1.02 BLEU 值。

系统	NIST02	NIST03	NIST04	NIST05	NIST08	平均
Baseline	37.38	35.00	38.04	34.32	25.74	34.09
BPE_cn	37.30	35.12	38.34	34.79	26.79	34.47
BPE_en	37.65	35.03	38.27	35.10	26.67	34.54
BPE_all	38.22ǂ	35.25ǂ	39.11ǂ	35.27ǂ	27.73ǂ	35.11

表 8：使用和未使用 BPE 的系统在测试集上的翻译性能 BLEU 值.

注：ǂ:表示在将显著水平设置为 0.01 时，BPE_all 系统比 Baseline 系统相比有显著性提高[10]。

5 总结

在本篇文章中，我们分析发现 BPE 编码的方式确实在一定程度上解决了 NMT 系统中未登录词的问题。通过将原有单词拆解为高频子字单元的方法，扩展了原有系统中的词汇表

[5] 本文的 SMT 系统采用 cdec 源码实现的层次短语翻译系统(https://github.com/redpony/cdec)。为公平起见，SMT 系统与 NMT 系统使用相同的训练集、开发集和测试集。

的大小，使在利用相同词汇表大小的情况下，我们可以表示出更多的单词，从而使系统中未登录词个数大大减少。

BPE 的方法和 SMT 系统相比，统计 UNK 单词被正确翻译的概率，我们又发现 BPE 方法在翻译精准度上基本和 SMT 系统持平，从而可以说明 BPE 方法在原有 NMT 系统流畅性的基础上又具有一定的翻译精准度。

但是使用 BPE 的方法仍然有其自身存在的问题，例如单词的漏翻现象。对于 NMT 系统中低频词和未登录词的问题仍然是一大难题。人们在人工智能的道路上依然任重道远。

参考文献

[1]N.Kalchbrenner and P.Blunsom.Recurrent continuous translation models.In *EMNLP*. 2013

[2]I.Sutskever,O.vinyalsand Q.V.Le.Sequence to sequence learning with neural networks. In *NIPS*. 2014

[3]Dzmitry Bahdanau, Kyunghyun Choand Yoshua Bengio.Neural MachineTranslation by Jointly Learning to Align and Translate. In *ICLR*. 2015

[4]Rico Sennrich and Barry Haddow.Neural Machine Translation of RareWords With Subword Units.In *ACL*. 2016

[5]Philip Gage .A New Algorithm for Data Compression. C User J.,12(2):23-38,February. 1994

[6]Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.Neural machine translation by jointly learning to align and translate. In *ICLR15*. 2015

[7]Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau,FethiBougares,Holger Schwenk,and Yoshua Bengio.Learning PhraseRepresentations using RNN Encoder-Decoder for Statistical Machine Translation.In *EMNLP*.2014

[8]Kishore Papineni,Salim Roukos,Todd Ward,and Wei jing Zhu. BLEU:amethod for automatic evaluation of machine translation .In *ACL*. 2002

[9]Mattew D.Zeiler. ADADELTA:An Adaptive Learning Rate Method.CoRR,abs/1212.5701. 2012

[10]Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*.2004

作者联系方式：韩冬，江苏省苏州市苏州大学本部理工楼 416 实验室，215006，18896556013,1196594306@qq.com

李军辉，江苏省苏州市苏州大学本部理工楼，215006, 13812696576, lijunhui26@gmail.com