神经机器翻译中英语单词及其大小写联合预测模型

张楠[1*]，李响[2,3]，靳晓宁[1]，陈伟[4]

（1. 北京工业大学 北京未来网络科技高精尖创新中心，北京 100124；

2. 中国科学院 计算技术研究所，北京 100190；

3. 中国科学院大学，北京 100049；

4. 北京搜狗科技发展有限公司，北京 100084）

摘要：英文中单词有大小写之分，如果使用不规范，会降低语句的可读性，甚至造成语义上的根本变化。当前的机器翻译处理流程一般先翻译生成小写的英文译文，再采用独立的大小写恢复工具进行还原，这种方式步骤繁琐且没有考虑上下文信息。另一种方式是抽取包含大小写的词表，但这种方式扩大了词表，增加了模型参数。本文提出了一种在神经机器翻译训练中联合预测英语单词及其大小写属性的方法，在同一个解码器输出层分别预测单词及其大小写属性，预测大小写时充分考虑源端语料和目标端语料上下文信息。该方法不仅减小了词表的大小和模型参数，翻译译文的质量也得到提升。在 WMT 2017 汉英新闻翻译任务测试集上，相比基线实验，我们提出的方法在大小写敏感和大小写不敏感两个评价指标上分别提高 0.97 BLEU 和 1.01 BLEU，改善了神经机器翻译模型的性能。

关键词：机器翻译；大小写恢复；联合预测

Joint Prediction Model of English Words and Their Cases in Neural Machine Translation

Nan Zhang[1]，Xiang Li[2,3]，Xiaoning Jin[1]，Wei Chen[4]

（1. Beijing Advanced Innovation Center for Future Internet Technology, Beijing University of Technology, Beijing 100124 China；

2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190 China；

3. University of Chinese Academy of Sciences, Beijing 100049 China;

4. Sogou Inc, Beijing, 100084 China）

Abstract: In English, words are case-sensitive. If the form of words is not standardized, it will reduce the readability of the article, and even cause a fundamental change in semantics. The current machine translation process is generally translated into lowercase English words, and then uses an independent case-recovery tool to restore. This method is cumbersome and does not consider contextual information. Another solution, the words list include uppercase and lowercase of words, but this approach expands the size of the words list and increases the number of parameters in the translate model. This paper proposes a method for jointly predicting English words and their case in neural machine translation. It predicts words and their case respectively in the same decoder, and fuse the information of the source corpus and the target corpus when predicting capitalization. The method not only decreases the size of the words list, reduces the parameters amount of the model, but also improves the quality of translation. Compared to the baseline system in the WMT 2017 Chinese-English news translation task test set , the proposed method improved 0.97 BLEU on case-insensitive, and improved 1.01 BLEU in case-insensitive.

Key words: Machine Translation; Case Restoration; Joint prediction

---

[*] 主要工作是在搜狗实习期间完成

0 引言

受现实应用的驱动，机器翻译近几年一直是备受关注的研究热点[1]。针对机器翻译，传统的解决方案是统计机器翻译。近几年深度学习在图像领域得到了很好的发展，在分类领域取得了超越人类的成绩[2]，受此影响，深度学习的方式也迅速在其他领域得到广泛应用。2014年，Jacob Devlin 提出了神经网络联合模型，相对于传统的统计机器翻译方法获得了显著的提升[3]。今年，微软 Hany 等人又应用神经机器翻译，将翻译的质量首次超越人类[4]。神经机器翻译逐渐成为机器翻译的主流方法，本文亦是采用神经机器翻译进行汉英翻译任务。

以往的汉英文翻译任务，生成的英文译文多为小写，需要额外的步骤恢复译文中单词的大小写信息。英语单词有大小写之分，一般情况下，单词的大小写形式可分为三种：全大写（USA、WTO 等）、首字母大写(China、Bill 等)、全小写（prediction、model 等）。同一单词的不同大小写形式，有时会代表不同的含义。比如“the white house” 可翻译为白色房子，但是“the White House” 则是特指“白宫”。不规范的书写形式，会极大的阻碍文本的可读性，降低阅读速度。当前很多机器翻译方法得到小写形式的英文译文后，通过使用大小写词表或者训练好的单词大小写恢复模型来恢复单词的原有大小写信息，增强译文的可读性。大小写恢复是对输入的单词序列，恢复其应有的大小写信息[5]。大小写恢复在命名体识别和语音识别等领域中亦有广泛应用[6][7]。

在这篇论文中，基于目前主流的 transformer 翻译模型[8]，我们提出了一种联合预测小写形式英文单词及其对应大小写属性的神经机器翻译方法，在同一个解码器输出层分别预测单词及其大小写属性。预测单词和预测单词大小写两项任务共享模型中的同一个解码器，在预测单词大小写属性时，不仅考虑了译文中单词的属性及位置，还充分融合了源端汉语的上下文信息。解码端预测单词及其对应单词的大小写是同时进行的，相较于传统方式减少了处理流程和处理时间。翻译预测结束后，根据解码得到的大小写类别信息，对小写译文中的单词进行大小写还原。在 WMT 2017 汉英新闻翻译任务测试集上，相比基线实验，我们提出的方法在大小写敏感和大小写不敏感两个评价指标上分别提高 0.97 BLEU 和 1.01 BLEU。

1 相关工作

针对恢复译文中英语单词的大小写，传统的处理方式主要有两种。一种是基于查表的方式，通过对训练语料中单词的各种大小写形式进行统计，将含有特定大小写信息的单词构建成一个表。在翻译得到译文后，译文中的每个单词根据词表选择一个可行性最大的形式进行恢复。该方法一般需要较大的词表才能达到一定的词语覆盖度。单词大小写形式与单词属性、在句子中所处的位置以及上下文语境都有关系，这种方式没有考虑译文的上下文信息，因此也容易产生错误恢复。而且在实际的数据中，同一单词可能有多种不同的大小写形式，会造成恢复结果的歧义。另一种译文大小写恢复的方法是训练一个单词大小写的恢复模型。比如，Lita 等人使用 trigram 模型恢复句子中的大小写信息[5]；Chelba 和 Acero 将大小写恢复视为一个序列标注问题[9]，并使用最大熵马尔科夫模型来融合单词和他们的大小写信息；Raymond 利用循环神经网络来在字符级别上预测单词大小写信息[10]。以上这些训练恢复模型的方法都是在单语料上进行。翻译结束后，针对目标端译文进行大小写恢复，增加了处理流程和时间开销。并且这些方法都没有考虑源端语料的情况，当翻译结果不准确时将对单词大小写信息的恢复产生极大干扰。

除了以上两种方式， Sennrich 和 Haddow 提出的 Byte Pair Encoding（BPE）[11]的方式也能在一定程度上解决译文大小写恢复的问题。BPE 将单词拆解为更小、更常见的子词单元。通过这种方式即在词表中保留字词的大小写属性，词表大小也未显著增大。

我们提出的联合预测模型，将预测单词和预测单词的大小写属性进行联合，在翻译预测

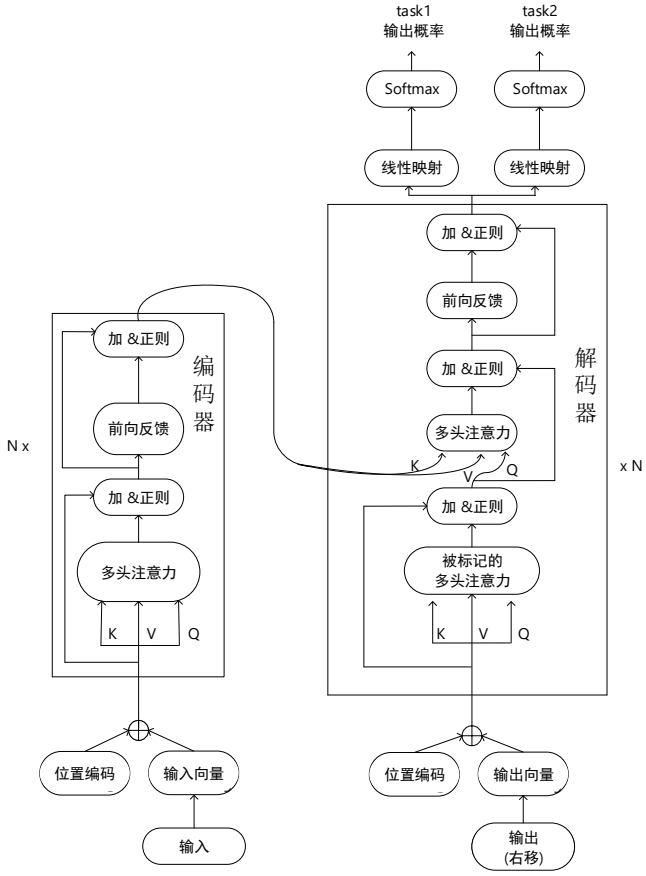
单词的同时，也预测单词的大小写属性。预测大小写时，不仅考虑了目标端英文单词的上下文信息，而且也考虑了源端汉语的上下文语义信息，因此能得到质量更好的译文。

2 联合预测模型

基于 transformer 模型的联合预测架构，将预测单词的任务$task_1$和预测单词大小写属性的任务$task_2$进行联合。

2.1 整体架构

在进行大小写预测时，用一个独立的解码器来预测单词的大小写属性，实际上会给模型增加很多的参数，加大了模型的训练难度及解码时间。单词的大小写形式很少，大致可以分为四类：全大写，开头大写，小写，其他类这四种情况。对于这种较少属性类别的预测，不需过多的参数，所以我们针对单词预测和大小写预测这两个任务采用共享解码器的方案。



图一、共享解码器联合预测架构

训练联合预测模型需要汉语语料、英语语料以及根据英语语料中单词原有大小写属性构建的英语单词标签语料。选取公开数据集中的汉英文平行语料，根据其中英语语料中单词的大小写属性构建英语标签语料。英语单词具有首字母大写、全大写、小写、其他四种大小写属性，根据英语语料中英文单词的大小属性，构建对应的单词属性训练语料。英语标签语料构建完成后，将英语语料中的单词全部转为小写。由此得到汉语语料、英语语料和英语标签语料。

基于 Transformer 模型，翻译模型由两部分组成：码器和解码器。编码器由一个多头注意力结构和一个前向反馈组成，解码器由两个多头注意力结构和一个前向反馈组成。多头注意力结构是用于学习单词或者词组之间的注意力，前向反馈学习语言内部的关系。将汉语语

料输入到编码器，经过多头注意力结构，编码器学习汉语词组之间的注意力，然后经过正则化处理做前向反馈，再经过正则处理输出到下一部分。此编码器处理过程重复 N 次。编码器每次正则化处理都要加上前一步的输入。编码器的输出即是解码器的部分输入。解码器的另一部分输入为英语语料。将英语语料输入到解码器时，英语词向量要右移一位。将输入的英语词向量序列通过做标记的方式，屏蔽还未翻译到的单词。然后解码器首先通过多头注意力结构学习英语单词之间的注意力，将结果正则化处理后与编码器的输出再次输入到一个多头注意力结构中学习汉语与英语之间的注意力，再将结果正则化处理后进行前向反馈，对前向反馈的结果再正则化处理后输入到下一部分。此解码器处理过程处理 N 次。解码器每次正则化处理也都要加上前一步的输入。有异于 Transformer 模型，本模型的解码器输出有两个预测任务，一个预测单词$task_1$，另一个用于预测单词的大小写信息$task_2$。解码器输出经过线性映射和 softmax 处理后预测单词，以英语词向量语料为真实标签求取预测损失。另一个解码器输出经过线性映射和 softmax 处理后预测单词大小写，以英语单词大小写标签为真实标签求取预测损失。所以，模型损失函数$Loss$由两部分组成，一部分是预测单词 task1 的损失，另一部分是预测单词大小写 task2 的损失。

$$Loss = loss_{task1} + \lambda loss_{task2} \tag{1}$$

两部分均使用交叉熵损失函数[12]。

2.2 点积注意力函数

图一中，模型注意力函数的输入 Q、K、V，分别代表 query、key-value 对。具体实现具体操作如图二（1）所示，根据 query 和 key 的相似度计算注意力权重。然后根据注意力权重，对 value 每个词向量进行加权即得到注意力。模型采用了缩放点积注意力（Scaled dot-product attention）：

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

其中 query Q 和 key K 的维度是相同的，都是d_k。Value V 的维度是d_v。其中标记（Mask）主要是用来去除矩阵乘后对角线之间的关系。

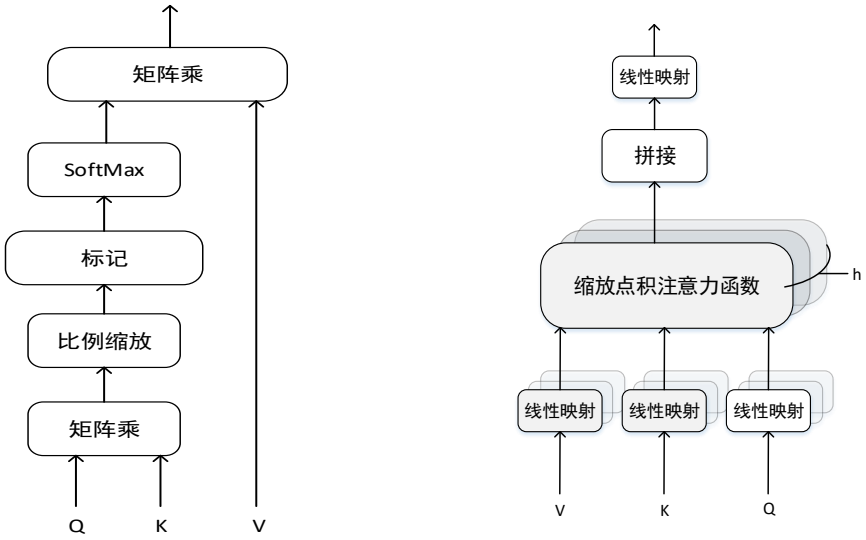
2.3 多头注意力的机制

模型采用了多头注意力的机制（Multi-Head Attention），如图二（2），将 Q、K、V 进行 h 次不同的线性映射，然后再将线性映射的结果映射到d_k，d_k，d_v维。分别对每一个映射之后的得到的 queries，keys 以及 values 进行注意力函数的并行操作，生成d_v维的输出值。

$$\text{MultiHead(Q,K,V)} = \text{Concat}(head_1, \dots, head_n)W^o \tag{3}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

其中$W_i^Q \in R^{d_{model} \times d_k}$，$W_i^k \in R^{d_{model} \times d_k}$，$W_i^V \in R^{d_{model} \times d_V}$，$W_i^O \in R^{d_{model} \times hd_v}$。$d_{model}$是模型的维度。



（1） （2）

图二：（1）点积注意力 （2）多头注意力

Transformer 模型没有使用循环神经网络或者卷积神经网络，为了使用序列的顺序信息，需要将序列的相对以及绝对位置信息加入到模型中去。因此，对汉英语料、标签语料分别抽取词表，建立单词与 ID 的映射，通过词表将语料转换成 ID 序列。再将汉、英序列以及标签序列转换成对应的词向量，在词向量中采用正弦和余弦函数[8]加入位置编码信息后输入到模型中。

3 实验及数据分析

3.1 实验数据

本实验采用的平行语料为 WMT2017 汉英任务训练数据中的 CWMT 部分数据，共 700 万条汉英数据。测试用的是 WMT2017 汉英新闻翻译任务测试集。

根据汉英训练语料里英语语料单词的大小写属性构建英语标签语料。我们将英语单词分为四种类别：a)其他，b)小写，c)开头大写，d)全大写。根据英语语料里英文单词的大小属性，构建对应的单词属性标签训练语料。英语单词标签训练语料构建完成后，将英语语料中的单词全部转换为小写形式。对于源端汉语语料，我们用 jieba 分词[†]将训练语句进行分词。至此得到了训练要用的汉语语料、英语语料以及英语标签语料。

对汉语语料、英语语料分别进行词频统计，取词频出现较高的单词构建汉语词表以及英语词表。针对训练语料中词表未覆盖到的单词，用 UNK 来表示。英语单词大小写的分类很少，所以选取全部的类别，即 a、b、c、d 四类，得到英语标签语料的标签词表。

3.2 实验设计

3.2.1 基本实验

我们设计了三组实验。如表所示

[†] https://github.com/fxsjy/jieba

表 1 实验词表大小

实验	源端词表大小	目标端词表大小
Baseline1	40K	60K
Baseline2	40K	93K
Our_Method	40K	60K

Baseline1：将训练数据和验证集中英语单词转为小写，抽取英语词表大小 6 万，词表对英文数据中单词的覆盖率达到 98%。汉语词表大小 4 万，对训练数据中分词后词组覆盖率达到 97%。同时验证集的英语端也转小写，用于测试，作为 Baseline1。

Baseline2：保留英语数据大小写信息，重新抽取英文词表，词表大小 9.3 万(与实验 1 英语词表的覆盖率保持一致)，汉语词表大小不变。

我们提出的联合预测方法:

Our_Method：根据单词所处的位置预测大小写信息。模型在预测单词的同时预测该单词可能的大小写信息。词表大小和 Baseline1 相同，汉语词表 4 万，英文词表 6 万。

3.2.2 BPE 实验

目前处理翻译译文大小写的主要方法是 Byte Pair Encoding（BPE）。BPE 方法将大小写敏感的语料拆解为常见的子词，在降低词表的同时又减少了译文中 UNK 的数量，从而极大地保存了句子的结构特征和流畅性。用 BPE 汉英平行语料进行处理。效果如下：

源端：企业 集团 就 网络 安全@@ 法 向 中国 提@@ 诉求 。

目标端：Business Groups Appeal to China Over Cyber@@ security Law.

BPE 将单词或词组拆解成了更小的组成部分。比如“安全法”拆解成“安全@@”和“法”，将“提速求”拆解成“提@@”和“诉求”，将“Cybersecurity”拆解成了“Cyber@@”和“security”

这个实验主要是用来验证联合预测的方式在 BPE 的方法下是否依然能取得较好的效果。根据 BPE 处理后的训练数据抽取词表。以 3.2.1 中 Baseline2 和 Our_Method 为基础设置对比实验 Baseline3 和 Our_Method_BPE。

Baseline3：除却词表大小和训练数据与 Baseline2 不同外，其余操作、设置均相同。

Our_Method_BPE：除却词表大小和训练数据与 Our_Method 不同外，其余操作、设置均相同。

表 2 BPE 实验词表达小

实验	源端词表大小	目标端词表大小
Baseline3	35K	35599
Our_Method_BPE	35K	29457

Baseline3 的目标端词表大小为 35599，Our_Method_BPE 的目标端词表大小为 29457. 两个词表对英文数据单词的覆盖度达到 100%。

在预测使用 beam search 解码时，大小写分类的选择并不参与 beam search，只是选取概率最大的一个类别作为预测单词大小写属性的结果。

我们在两张 Titan XP 训练我们的模型。在 tensor2tensor 框架[13]下，基于 transformer 模型实现程序。Transformer 中$N = 4$，4 个编码层 4 个解码层，词向量（embeding）为 512 维度，隐层的维度是 1024。Batch 大小为 4096，学习率 0.1，warm up 为 4000。损失函数中$\lambda = 1$。其他参数选用的均是 transformer_base 的参数。

3.3 实验结果

我们使用机器翻译领域常用的 BLEU[14]作为评价指标来比较各个实验的结果，脚本使用 Moses 系统[15]提供的 multi-bleu.pl[‡]。

3.3.1 基本实验

表 2 基本实验结果

实验	目标端英语大小写敏感	目标端英语大小写不敏感
Baseline1		18.29
Baseline2	18.22	19.00
Our_Method	19.19	20.01

由上表可知，我们的方法在大小写敏感和不敏感的两个指标上均高于 baseline2，高出 baseline2 一个 BLEU 左右。大小写不敏感也高于 baseline1 联合的方式.不仅在翻译的同时预测单词大小写，同时还提升了译文的质量。

由于三个实验的词表大小有所不同，我们还统计了四个实验结果中 UNK 字符的数量。

表 3 实验结果 UNK 数量

实验	UNK 数量
Baseline1	8306
Baseline2	1801
Our_Method	1782

由表 3 可知，Baseline2 和 Our_Method 的 UNK 均比 Baseline1 少。Baseline2 的目标端英文词表（9.3 万）比 Baseline1 的词表（6W）要大，所以降低了译文中的 UNK 数量。Baseline1 和 Our_Method 的英文词表虽然相同，但是由于 Our_Method 同时预测了单词的大小写信息，所以 Our_Method 的英文词表的可表示单词量远大于 Baseline1，以此降低了 UNK 的数量。

由于 Baseline1 的 UNK 数量非常多。在去除结果中所有的 UNK 后，再次测试了 BLEU 结果。

表 4 去除 UNK 后结果

实验	目标端英语大小写敏感	目标端英语大小写不敏感
Baseline1		**19.45**
Baseline2	18.22	19.00
Our_Method	19.20	20.01

由表 4 可知，在排除 UNK 影响后，Baseline1 的大小写不敏感结果要优于 Baseline2。虽然 Baseline2 的英文词表（9.3 万）与 Baseline1 的词表对训练语料具有相同的单词覆盖度，但是词表的增大也增加了模型的训练参数，提升了模型的训练难度，进而影响译文质量。由于 Baseline2 和 Our_Method 的 UNK 数量较少，所以去除 UNK 后的结果基本没有变化。Our_Method 结果依然比两个 Baseline 的翻译质量要好。与 Baseline1 相比，两者具有相同的词表大小，但是 Our_Method 由于预测了大小写属性，增加了可表示单词的数量，扩大了单

[‡] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

词覆盖率。从图二联合预测的翻译架构可知，模型的学习的注意力分为三部分：源端汉语注意力、目标端英语注意力以及源端汉语和目标端英语之间的注意力。当扩大了英文词表的大小或者提升了词表的可表示单词数量，可以让模型学习到更多英语单词之间的注意力分配机制以及汉语字词与英语单词之间的注意力关系，使模型在翻译预测单词时能够做出更精确的预测。Our_Method 与 Baseline2 相比，虽然汉语词表大小相同且英语词表对训练数据具有相同的覆盖度，但是由于 Our_Method 预测单词有四种分类，所以实际可表示的单词数量比 Baseline2 的词表要多。另一方面，Our_Method 的词表大小比 Baseline2 小了 3.3 万，这也减少了模型的参数，更有利于模型训练。

3.3.2 BPE 实验

表 5 BPE 实验结果

实验	目标端英语大小写敏感	目标端英语大小写不敏感
Baseline3	20.21	21.24
Our_Method_BPE	20.43	21.49

从表 5 可知多任务联合预测的 Our_Method_BPE 结果要好于 Baseline3 的结果，但是不像 Our_Method 与 Baseline2 相比提升的那么明显。同时，通过比较 Baseline2 和 Baseline3，我们可知 BPE 处理数据后训练出的模型，性能也较优。通过统计翻译结果，译文中未发现 UNK，这是由于 BPE 通过分解子词的方式，有效的提升了字词词表对训练数据的覆盖度，英语词表对训练数据的覆盖度达到了 100%，所以在结果中没有出现 UNK 的情况。在英语词表对训练数据的覆盖度达到了 100%的情况下，通过预测子词的大小写属性，增加的可表示单词数量有限。同时联合预测方式，英文词表比 Baseline3 英文词表小了 6142，在一定程度上有所减小，所以实验(2)的结果对 Baseline3 有所提升，但是提升不像之前实验那么明显。

4 总结

本文以汉英翻译中英文单词的大小写预测为研究对象，提出了一种在神经机器翻译训练中联合预测英语单词及其大小写属性的方法。以往的大小写恢复多是在机器翻译结束后，根据译文恢复单词的大小写信息。本文提出的方法，综合考虑了源端和目标端两者的信息，根据单词所处的位置以及单词本身的属性预测，达到了很高的准确度。由于联合预测大小写的方式降低了词表的大小并且提升了词表的可表示单词数量，使模型可以学习到更多单词之间的注意力关系，降低模型参数数量的同时还提升了翻译译文的质量。在 WMT 2017 汉英新闻翻译任务测试集上，本文提出的联合预测的方法在大小写敏感和不敏感两个指标上均高于 Baseline。

参考文献

[1] 李茂西, 宗成庆. 机器翻译系统融合技术综述[J]. 中文信息学报, 2010, 24(4):74-85..

[2] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition[C], Proceedings of the CVPR, 2016: 770–778.

[3] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, Fast and Robust Neural Network Joint Models for Statistical Machine Translation[C], 2014,Proceedings of the 52nd An- nualMeeting of the Association for Computational Lin- guistics, vol. 1, pp. 1370–1380.

[4] Hassan H, Aue A, Chen C, et al, Achieving Human Parity on Automatic Chinese to English News

Translation [J], 2018. arXiv PrePrint arXiv:1803.05567

[5] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla, tRuEcasIng[C],2003 , Proceedings of the ACL 2003: 152–159.

[6] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1):p ágs. 3-26.

[7] Batista F, Moniz H, Trancoso I, et al. Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(2):474-485.

[8] Vaswani A, Shazeer N, Parmar N, et al, Attention Is All You Need[C], 2017,Proceedings of the NIPS.

[9] Chelba C, Acero A. Adaptation of maximum entropy capitalizer: Little data can help a lot ☆[J]. Computer Speech & Language, 2006, 20(4):382-399.

[10] R. H. Susanto, H. L. Chieu, and W. Lu, Learning to Capitalize with Character-Level Recurrent Neural Networks: An Empirical Study[C],2016, Proceedings of the Emnlp, vol. 5, no. 3:. 128–137.

[11] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.

[12] Theodoridis S. Chapter 18 - Neural Networks and Deep Learning[M]// Machine Learning. Elsevier Ltd, 2015:875-936.

[13] Kaiser L, Gomez A N, Shazeer N, et al. One Model To Learn Them All[J]. 2017.

[14] Kishore Papineni, Salim Roukos, Todd Ward,et alBLEU: a Method for Automatic Evaluation of Machine Translation[C], 2002,Proceedings of the ACL, 311-318

[15] Koehn, Philipp, Hoang, et al. Moses: open source toolkit for statistical machine translation[C]// Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007:177-180.

作者联系方式：张楠 北京市朝阳区平乐园 100 号北京工业大学 100124 18811715993 zhangnan3065@foxmail.com