doi:10.6043/j.issn.0438-0479.201811023

融入汉字笔画序列的神经机器翻译

谭 新,邝少辉,张龙印,熊德意*
(苏州大学计算机科学与技术学院,江苏 苏州 215006)

摘要:神经机器翻译(NMT)因其在多个语言对上的翻译效果都远超传统的统计机器翻译(SMT)而逐渐成为机器翻译方向的主流.然而,这种 NMT 系统在将向量化的词语作为输入时只考虑了词语整体的语义信息,忽略了构成词语的汉字本身所包含的信息.为此,针对汉字给出了一种融入汉字笔画序列的 NMT 系统.该系统在将词语的词向量作为输入的同时又将向量化的汉字笔画序列作为额外输入,既考虑了中文词语整体的语义信息,又考虑了构成词语的汉字本身的内部语义信息和外部形态信息.实验结果表明,提出的融入了汉字笔画序列的 NMT 系统更加有效,其翻译结果更加准确流畅,与传统的 NMT 系统相比机器双语互译评估(BLEU)值能够提高 1.21 个百分点.

关键词:神经机器翻译;汉字笔画序列;注意力机制
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2019)02-0164-06

机器翻译是利用计算机算法自动学习并主动将一种源语言翻译成另一种目标语言的过程,隶属自然语言处理领域,具有重要的科研意义和实用价值.近年来,Sutskever 等[1]和 Cho 等[2-3]提出的神经机器翻译(NMT)系统因其显著提高的流畅度以及在多个语言对上远超传统的统计机器翻译(SMT)[4]系统的翻译效果,逐渐代替 SMT 系统成为机器翻译方向的主流.然而,这种 NMT 系统虽然将词语的潜在语义空间向量表示作为输入,考虑了词语级别的整体语义信息,却忽略了构成词语的子字级别信息.

针对这一问题,也有一些考虑子字级别信息的方法,如:Cotterell 等[5]将词的形态信息融入词向量的学习过程;Bojanowski 等[6]将字符的 n-元信息融入词向量的学习过程.然而,这些方法都是针对欧洲语言如英语、德语、西班牙语等由拉丁字母构成的书写系统,并不适用于中文汉字这种由笔画构成的书写系统.对于汉字的书写特性,每个中文词语由少量的汉字构成,且每个汉字由一系列的笔画序列组成,因此每个汉字本身也表达了丰富的信息,包括内部语义信息和外部形态信息.Kuang 等[7]曾提出利用抽取的汉字偏旁辅助 NMT 的方法,考虑一些由偏旁相同引起汉字语义相近的情况,如“林”、“树”等字都是“木”字旁,表达含义都和“树木”有关,因此可以使 NMT 系统学到这类偏旁包含的语义信息;但是用这种方法得到的信息既不完整又容易引入大量噪声,如“智”、“晶”这类汉字,虽然偏旁都为“日”,但是“智”表达“智慧、知识”,“晶”却表达“晶莹、明亮”,两者表达的语义完全不同.

为了更有效地捕获构成词语的汉字本身所携带的信息,本文中考虑每个汉字的完整笔画级别信息,提出了一种将汉字构字的笔画序列融入 NMT 系统的方法,以期捕获汉字的语义和形态信息,使译文更加准确、流畅.

1 带有注意力机制的 NMT 系统

带有注意力机制的 NMT 系统[8]通常采用编码器-解码器(encoder-decoder)的架构.其中,编码阶段将整个句子作为输入进行编码并输出包含句子信息的向量;解码阶段将编码阶段得到的向量作为输入,逐次生成目标语言的单词序列.下面对编码器和解码器进行详细说明.

编码器由双向循环神经网络(Bi-RNN)构成,首先将构成句子的词语映射为词向量序列,即 $X \rightarrow [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{T_X}]$,其中 X 为输入的句子,$\boldsymbol{w}_m$ 是一个多维

收稿日期:2018-11-12 **录用日期**:2018-12-05
基金项目:国家自然科学基金(61622209,61861130364)
* **通信作者**:dyxiong@suda.edu.cn
引文格式:谭新,邝少辉,张龙印,等.融入汉字笔画序列的神经机器翻译[J].厦门大学学报(自然科学版),2019,58(2):164-169.
Citation: TAN X, KUANG S H, ZHANG L Y, et al. Integration of Chinese character stroke sequence into neural machine translation[J]. J Xiamen Univ Nat Sci, 2019, 58(2): 164-169. (in Chinese)

的向量,表示源端句子中第 m 个词 x_m 的词向量.前向 RNN 根据公式 $\overrightarrow{\boldsymbol{h}}_m = f(\boldsymbol{w}_m, \overrightarrow{\boldsymbol{h}}_{m-1})$ 获得由隐层向量组成的前向向量序列 $[\overrightarrow{\boldsymbol{h}}_1, \overrightarrow{\boldsymbol{h}}_2, \cdots, \overrightarrow{\boldsymbol{h}}_{T_X}]$;反向 RNN 依据同样的原理得到由隐层向量组成的反向向量序列 $[\overleftarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_2, \cdots, \overleftarrow{\boldsymbol{h}}_{T_X}]$.然后将 $\overrightarrow{\boldsymbol{h}}_m$ 和 $\overleftarrow{\boldsymbol{h}}_m$ 进行拼接得到 $\boldsymbol{h}_m$,即 $\boldsymbol{h}_m = [\overrightarrow{\boldsymbol{h}}_m; \overleftarrow{\boldsymbol{h}}_m]$,作为含有上下文信息的隐层向量序列表示 $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{T_X}]$.

在此之后,隐层向量序列借助注意力网络获得上下文向量 $\boldsymbol{c}_t$,具体运算公式如下:

$$\boldsymbol{c}_t = \sum_{j=1}^{T_X} \alpha_{tj} \boldsymbol{h}_j, \tag{1}$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T_X} \exp(e_{tk})}, \tag{2}$$

$$e_{tj} = a(\boldsymbol{s}_{t-1}, \boldsymbol{h}_j). \tag{3}$$

其中:α_{tj} 是编码器中 t 时刻第 j 个编码端隐层状态 $\boldsymbol{h}_j$ 的权重;解码器由 RNN 构成,e_{tj} 是一个对齐模型,用来计算 t 时刻生成的目标端句子词语与第 j 个源端句子词语的匹配程度;a 是一个一层的前向网络;$\boldsymbol{s}_{t-1}$ 为 $t-1$ 时刻解码端的 RNN 隐层状态.

在解码阶段,对于给定的 context 向量 $\boldsymbol{c}_t$ 以及所有已经预测得到的目标端词集 $\{y_1, y_2, \cdots, y_{t-1}\}$,通过以下公式来预测每个词 y_t:

$$p(Y) = \prod_{t=1}^{T_{y_t}} p(y_t \mid \{y_1, y_2, \cdots, y_{t-1}\}, X), \tag{4}$$

$$p(y_t \mid \{y_1, y_2, \cdots, y_{t-1}\}, X) = g(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t). \tag{5}$$

其中:$Y = (y_1, y_2, \cdots, y_{T_{y_t}})$,表示目标端句子;$X$ 为给定的输入序列;g 是非线性激活函数,通常使用 softmax 函数;$\boldsymbol{s}_t = f(y_{t-1}, \boldsymbol{s}_{t-1}, \boldsymbol{c}_t)$.

最后,系统采用随机梯度下降的方法对目标函数进行训练迭代,目标函数如下:

$$L(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_{y_t}} \log\, p(y_t^n \mid y_{<t}^n, x^n), \tag{6}$$

其中 x^n, y^n 表示训练集上的平行语句对.

2　融入汉字笔画序列的改进系统

2.1　中文汉字的笔画序列

英语、德语等欧洲语言的词语是由不包含语义信息的拉丁字母首先组成若干个包含语义信息的词根、词缀,随后将这些词根、词缀进行组合拼接构成.不同于欧洲语言,中文词语由多个汉字组合而成,而每个汉字又由包含语义信息的若干笔画序列组成.一个汉字往往包含丰富的内部语义信息和外部形态信息,然而这些信息不仅来源于一个汉字的偏旁或部首(像上文提到过的"智"和"晶",虽然含有相同的偏旁,但是表达含义却不同),更来自笔画序列的整体.因此,不希望神经网络学到不完整的甚至是包含大量噪声的信息,而是能够自动学习给定汉字的完整笔画序列所涵盖的语义和形态信息,进而推断由相关笔画序列构成的汉字所涵盖的信息,即像人类学习汉字一样,在学到汉字"水"和"冷"后,能够根据"水"的整体笔画序列和"冷"的偏旁部分("冫")的笔画序列推测出汉字"冰"的含义,如图 1 所示.

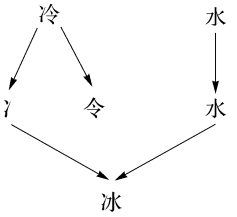

图 1　由已知汉字的笔画推测相关汉字含义的示例

Fig. 1　Example of speculation on the meaning of related Chinese characters from strokes of known Chinese characters

Cao 等[9]也曾利用这类汉字笔画序列提出了基于汉字笔画的中文词向量表示,有效证明了汉字笔画序列的可用性和重要意义.

结合以上两点,可以认为将这类汉字子字级别的笔画序列融入 NMT 系统能够有效提高翻译质量.在表 1 所示的笔画序号表中列出了国家规定的 5 种基本笔画,并将它们分别对应于笔画数字 1～5.

表 1　笔画序号

Tab. 1　Stroke serial number

笔画名	笔画数字
横(提)	1
竖(竖钩)	2
撇(竖撇)	3
点(捺,提捺)	4
折	5

由于汉字的书写过程有着严格的笔画先后顺序,所以对于每一个汉字,都可以根据它的笔画书写顺序得到唯一的笔画序列.对该笔画序列中的笔画依次用笔画序号表(表 1)中与之对应的笔画数字表示,将得到的数字序列称为笔画数字序列.例如:根据汉字"大"的笔画书写顺序将其拆分得到笔画序列"横('一')、撇('丿')、捺('㇏')",再借助笔画序号表(表

1)中的映射关系将该笔画序列转换为对应的笔画数字序列“134”.

2.2 融入汉字笔画序列的 NMT 系统

如前文所述,由于汉字笔画序列包含丰富的内部语义信息和外部形态信息,所以本研究设法将这些笔画序列作为额外信息融入 NMT 系统中,使得改进后的 NMT 系统能够从中文词语中学习到更多的有用信息,从而提高句子的翻译质量. 图 2 即为提出的融入汉字笔画序列的 NMT 系统.

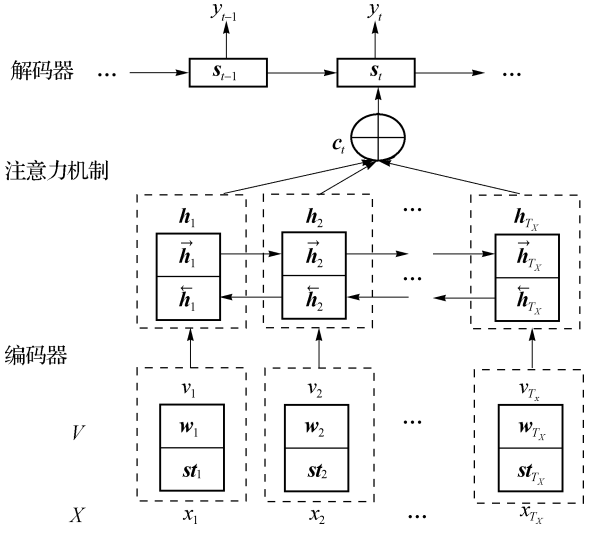

图 2 融入汉字笔画序列的 NMT 系统

Fig. 2 The NMT system with Chinese character stroke sequence

在提出的改进 NMT 系统中,输入序列 X 中包含 T 组词语$\{x_1,x_2,\cdots,x_{T_X}\}$,用词向量和笔画序列向量两部分表示该序列的潜在语义空间. 首先,用随机初始化的词向量对输入序列 X 进行如下表示:

$$\text{embedding\_seq}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{T_X}\}. \tag{7}$$

对于由 M 个汉字组成的第 i 个词 x_i,每个汉字由 N 个笔画构成,每个笔画来源于笔画序号表(表 1),用 dict_{mn} 表示第 m 个字的第 n 个笔画在笔画序号表中对应的笔画数字. 用如下的形式表示词 x_i 对应的笔画数字序列:

$$\text{stroke}_i=(\text{dict}_{mn}\mid m\in[1,M],n\in[1,N]). \tag{8}$$

为了充分利用神经网络模型的学习能力,将笔画数字序列 stroke_i 中的每个笔画数字 k 转换成对应的空间向量表示 $\textbf{vector}_k$,形成$(k,\textbf{vector}_k)$对,从而将第 i 个词的笔画数字序列映射为一组空间向量 $\textbf{vector}_{\text{stroke}_i}$. 经过实验发现,将每组词的笔画序列向量求平均能得到较好的效果,最终将 x_i 的笔画序列映射为如下的潜在语义空间表示:

$$\boldsymbol{st}_i=\text{average}\Big(\sum_{k\in\text{stroke}_i}\textbf{vector}_k\Big). \tag{9}$$

结合以上描述,将序列中词语的词向量表示和含有笔画序列的向量表示进行拼接,最终得到如下所示的输入端表示:

$$V=\{v_1,v_2,\cdots,v_{T_X}\}=\{[\boldsymbol{w}_i,\boldsymbol{st}_i]\mid i\in[1,T_X]\}. \tag{10}$$

将构成词语的汉字笔画向量 $\textbf{vector}_k$ 和词向量 $\boldsymbol{w}_i$ 分别固定维度并在开始时随机初始化,所有权重值随整个训练过程的进行而迭代更新,最终将笔画序列信息融入潜在语义空间以丰富词语的表示.

3 数据来源与参数设置

3.1 实验数据

使用语言数据联盟(Linguistic Data Consortium, LDC;www.ldc.upenn.edu)中抽取的 1.25×10^6 句中英平行句对作为训练集对 NMT 模型进行训练,其中包含了 8.09×10^7 个中文词语和 8.64×10^7 个英文单词. 另外,将国家标准与技术研究所(NIST)的 NIST06 数据集作为开发集,NIST 02,03,04,05,08 数据集(每个数据集包含 4 个目标译文)作为测试集进行模型评测,具体的实验相关数据的统计情况如表 2 所示.

表 2 实验相关数据统计

Tab. 2 Statistics of experimental related data

数据集	句子数	源端单词数
NIST 02	878	22 639
NIST 03	919	24 095
NIST 04	1 788	49 438
NIST 05	1 082	29 897
NIST 06	1 664	38 349
NIST 08	1 357	32 310

3.2 实验设置

对于实验中的基线(Baseline)系统利用 RNNSearch 实现. 为了有效训练模型,设置最大句子长度为 50 个词,同时将源端和目标端的词语表大小设定为出现频率最高的前 3×10^4 个词,其中各包含一个特殊标记“UNK”来表示未登录词,词表分别覆盖源端训练语料 97.7% 的中文词语和目标端训练语料 99.3% 的英语单词,词向量维度设置为 620 维. 实验中设置的编码

器和解码器的隐层数量各为 1,隐层单元个数各为 1 000,输出层使用 dropout 策略[10]和 Adadelta 算法[11]对神经网络中的参数进行更新,其中 dropout 设为 0.5.此外,batch_size 设为 80,并额外采用 beam_size 为 10 的集束搜索(beam search)算法寻找最可能的翻译结果.采用在开发集上获得最高机器双语互译评估(BLEU)值[12]评分的模型作为最终模型对测试集进行评测.

针对提出的融入汉字笔画序列的 NMT 系统,仍然使用基于注意力机制的 NMT 系统.其中,汉字笔画序列对应的笔画数字序列来自新华字典(http://xh.5156edu.com/),除此之外的其他实验设置均和上述的 Baseline 系统保持一致.对于融入的汉字笔画序列向量,分别将其向量维度设置为 40,100 和 200 维进行 BLEU 值对比,实验发现在汉字笔画序列的向量维度设置为 100 维时获得的 BLEU 值评分最高,翻译效果提升也最显著.

4　实验结果和分析

4.1　主要实验结果

表 3 给出了提出的融入汉字笔画序列的 Stroke_seq 方法与 Baseline 方法在开发集和测试集上的 BLEU 值对比结果.

表 3　Stroke_seq 与 Baseline 方法的 BLEU 值对比

Tab.3　Comparison of BLEU values between Stroke_seq and Baseline methods　%

方法	BLEU						
	开发集	测试集					
	NIST 06	NIST 02	NIST 03	NIST 04	NIST 05	NIST 08	均值
Baseline	34.78	38.99	36.90	39.18	36.31	26.21	35.52
Stroke_seq	36.18	39.53	38.25	41.02	37.08	27.79	36.73

Baseline 方法是使用 RNNSearch 实现的基于注意力机制的 NMT 系统;Stroke_seq 方法是使用本文中提出的融入汉字笔画序列的 NMT 系统,其使用笔画序列向量维度为 100 维.从表 3 可以看出 Stroke_seq 方法能够显著提高翻译的准确性,平均 BLEU 值超过 Baseline 方法 1.21 个百分点.

4.2　Stroke_seq 与 Radicals 结果对比

Radicals 方法使用 Shao 等[7]提出的利用抽取出的汉字偏旁辅助 NMT 系统,由于其仅在 NIST 06 和 08 上进行了实验,所以只对这两个测试集上的 BLEU 值进行对比,如表 4 所示.

表 4　Stroke_seq 与 Radicals 方法的 BLEU 值对比

Tab.4　Comparison of BLEU values between Stroke_seq and Radicals methods　%

方法	BLEU	
	NIST 06	NIST 08
Radicals	35.62	26.55
Stroke_seq	36.18	27.79

从表 4 可以看出,融入完整笔画序列的 Stroke_seq 方法明显优于仅利用汉字偏旁信息辅助 NMT 系统的 Radicals 方法,说明本文中提出的方法能够使神经网络主动学习到构成词语的汉字本身携带的信息,有效避免了由仅考虑汉字偏旁的片面信息所带来的噪声.

4.3　翻译示例

从测试集中随机抽取两个句子,将本文中提出的融入汉字笔画序列的 Stroke_seq 方法与 Baseline 方法生成的译文进行对比,结果如表 5 所示,其中,“Ref_”为原文句子对应的翻译参考译文,“Baseline_”为 RNNsearch 生成的译文,“Stroke_seq_”为提出的融入汉字笔画序列的 NMT 结果.从第一个示例可以看出,融入汉字笔画序列的 NMT 系统的翻译结果更加贴近参考译文,翻译流畅度更高;然而在第二个示例中,虽然 Stroke_seq 方法与 Baseline 方法翻译质量相比有很大提升,但同时也揭露了该方法仍存在翻译不充分的问题.

5　结　论

由于汉字的构成不同于欧洲语言的词根、词缀构词法,所以为了使翻译系统不仅考虑中文词语级别的整体语义信息,更能考虑到构成词语的汉字子字级别

表 5 翻译示例对比

Tab. 5 Comparison of translation examples

原文_1	诺 桑 比亚 警 局 也 表示 ，没有 证据 显示 这 两 起 窃案 和 史蒂 文生 爵士 的 工作 有关.
Ref_1	the northumbria police bureau also stated that there is no evidence showing that the two larceny cases are related to the work of sir stevens.
Baseline_1	UNK UNK said that there was no evidence to show that the two burglary cases were related to sir UNK's work.
Stroke_seq_1	the UNK police bureau also said that there was no evidence that these two burglary and UNK were related to the work of the UNK.
原文_2	伦敦 每 日 快报 指出 ，两 台 记载 黛 安娜 王妃 一九九七年 巴黎 死亡 车祸 调查 资料 的 手提 电脑 ，被 从 前 大 都会 警察 总长 的 办公室 里 偷 走.
Ref_2	London daily express pointed out that two laptop computers, which contain investigation materials about the car accident and death of princess Diana in 1997, were stolen from the office of the former metropolitan police chief.
Baseline_2	in London, the daily express said that the mobile computers of the UNK UNK UNK UNK UNK were found in the office of the former UNK UNK.
Stroke_seq_2	the London daily pointed out that the two records of the two UNK were stolen from the office of the chief superintendent of the city of UNK in the past.

的内部语义信息和外部形态信息，本文中提出了一种将带有丰富信息的汉字笔画序列融入基于注意力机制的 NMT 系统的方法. 首先将构成词语的带有丰富信息的汉字笔画序列转化为潜在语义空间向量表示，随后将词语的这种汉字笔画序列向量表示作为额外信息和原有的词向量拼接起来作为编码器的输入. 经实验验证，该方法能够显著提高中文到英文的翻译准确性和译文的流畅度.

尽管这种融入汉字笔画序列的 NMT 系统在处理源端语言为中文的翻译上颇为有效，但是处理的语种范围过于局限. 因此，下一步考虑将这种融入子字级别信息的方法应用于其他由笔画构成书写的语种，并探索更加通用的融合子字级别信息的方法以辅助 NMT 系统.

参考文献：

[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. Montreal: NIPS, 2014: 3104-3112.

[2] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1406.1078.

[3] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: encoder-decoder approaches [EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1409.1259.

[4] CHIANG D. A hierarchical phrase-based model for statistical machine translation[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Arbor: ACL, 2005: 263-270.

[5] COTTERELL R, SCHÜTZE H. Morphological word-embeddings[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Colorado: NAACL, 2015: 1287-1292.

[6] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1607.04606.

[7] KUANG S, HAN L. Apply Chinese radicals into neural machine translation: deeper than character level [EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1805.01565.

[8] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1409.0473.

[9] CAO S, LU W, ZHOU J, et al. Cw2vec: learning Chinese word embedding with stroke *n*-gram information[C]// The 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 5053-5061.

[10] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[11] ZEILER M D. ADADELTA: an adaptive learning rate method[EB/OL]. [2018-11-11]. https://arxiv.org/pdf/1212.5701.

[12] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a

method for automatic evaluation of machine translation [C] // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Pennsylvania: ACL,2002:311-318.

Integration of Chinese character stroke sequence into neural machine translation

TAN Xin, KUANG Shaohui, ZHANG Longyin, XIONG Deyi*

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

Abstract: Neural machine translation (NMT) has gradually become the mainstream of machine translation because it surpasses statistical machine translation (SMT) in a variety of language pairs. However, conventional NMT systems merely take word embeddings into account, ignoring the information contained in Chinese characters that constitute words. Therefore, this paper proposes a new approach that integrates Chinese character stroke sequence information into NMT system. Our proposed method considers both word embeddings for Chinese words and Chinese character stroke sequence information into NMT system. We consider not only the semantic information of each Chinese word, but also the internal semantic information and external form information of Chinese characters that form these words. Experimental results show that our proposed method, which integrates Chinese character stroke sequence information into NMT system, is more effective and makes translation results more accurate. Compared with traditional NMT system, our proposed method can achieve an improvement of BLEU value by 1.21 percentage points.

Keywords: neural machine translation; Chinese character stroke sequence; attention mechanism

论文撤回声明

发表在《厦门大学学报(自然科学版)》2018 年第 6 期 890—895 页的论文《基于门限卷积神经网络和词嵌入的中文分词法》由于第一作者张竞对公式理解得不够深入,造成公式的使用错误而得出错误的数据及结论,具体公式为文中 2.5 节公式(4)～(6).现本人提出撤稿申请,经编辑部核查,为了避免误导广大读者,同意其撤回此文.

2019 年 2 月 20 日

《厦门大学学报(自然科学版)》编辑部