BERT 符合中文分词

杨海琴

[1]美图秀秀

[2]香港恒生大学计算机系

haiqin.yang@gmail.com

ar Xiv: 1909.09292v1[cs。 CL]2019 年

摘要。 汉语分词是汉语理解的一项基本任务。 最近，基于神经网络的模型在解决域内 CWS 任务方面取得了优异的性能。 去年，变形金刚双向编码器表示(BERT)作为一种新的语言表示模型，被提出为许多自然语言任务的骨干模型，并重新定义了相应的性能。 BERT 的优异性能促使我们应用它来解决 CWS 任务。 通过对第二次国际汉语分词 Bake-off 的基准数据集进行密集的实验，我们得到了几个敏锐的观察结果。 即使数据集包含标记不一致的问题，BERT 也能稍微提高性能。 当应用足够学习的特征时，Softmax，一个更简单的分类器，可以获得与更复杂的分类器相同的性能，例如条件随机场(CRF)。 误码率性能通常随着模型尺寸的增加而增加。 通过 BERT 提取的特征也可以作为其他神经网络模型的良好候选。

1 导言

汉语分词(CWS)，即将文本划分为单词，是汉语理解[19]的关键预处理步骤。 此任务可以建模为令牌标记任务或基于字符的序列标记任务[10]。

最近，神经网络模型被应用于解决这一任务，在特征工程[1,4,10,11]中花费的精力较少。 例如，[11]，Max-MarginTensor 神经网络(MMTNN)被提出来建模标签和上下文字符之间的交互作用。 [3]，利用门控递归神经网络(GRNN)对 CWS 的字符组合进行建模。 [4]，提出并评价了四种不同的长时记忆(LSTM)体系结构，以测试 CWS 的性能。 [15]，卷积神经网络与 CWS 的单词嵌入相结合。 [10]介绍了一个针对 CWS 的 LSTM 的深入研究。 这些方法的本质归结为两个问题：1)如何有效地表示每个字符？ 2)如何吸收字符之间的转换来利用上下文信息？

为了解决上述问题,Yang 等人。从丰富的外部资源中学习了字符、字符 Bigram 和单词的预先训练的字符/单词嵌入，并在 CWS[18]中显示出显著的错误减少。 grnn，

源标签:	s	s	b e	b e	s
资料来源:	汇合点	于	2004 年	首发	。
含义:	合流于 2004 年首次释放。				
误码率标签:	b m	我		s	是的 s
伯特:	康#fl	#ue#nce		于	2004 年首发

无花果。 1. 中文单词标签的一个例子：一个轻微的区别在于源标签和 BERT 标签用于处理英语单词，请参阅文本中的详细描述。

应用 LSTM 和 CNN 对分段句子[3,4,15]的一致性进行了建模，但它们需要指定一个固定的上下文窗口，这缺乏充分捕捉上下文信息的灵活性。 在[1]中，采用门控组合神经网络对字符进行字表示生成，并采用 LSTM 评分模型进行分割，克服了固定大小上下文 windown 的局限性。 通过贪婪的搜索[2]进一步加速了分词。 然而，这些方法没有开发足够的外域资源，可能限制提高性能的潜在力量。

目前，大量外域资源的无监督学习的巨大语言模型，如 ELMO[12]和开放 AIGPT[13]，已经证明了利用从外域资源中学习的信息的前景。 特别是从变形金刚(BERT)[6]提出了双向编码器表示，并重新定义了十一种自然语言处理任务的技术状态。 BERT 的优异性能及其在文本中捕获上下文的能力促使我们将其作为预处理步骤来为 CWS 提取特征。

在本文中，我们试图了解在解决 CWS 任务的各个方面可以获得什么性能 BERT：

BERT 能继续提高 CWS 任务的性能吗？

— 字符表示和分类器之间的权衡是什么？

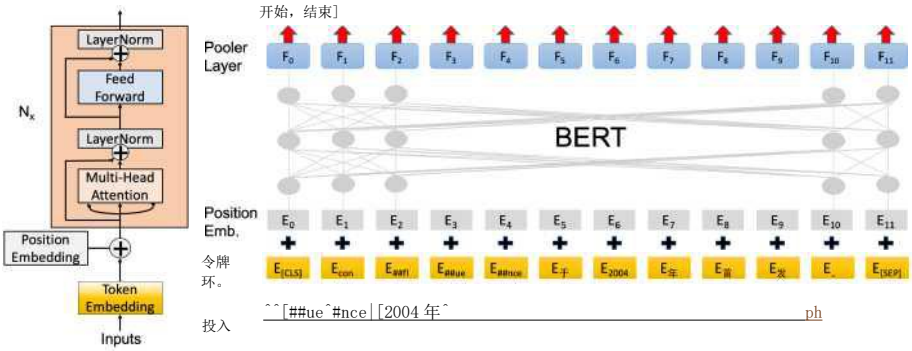
BERT 模型大小对 CWS 任务有什么影响？

— 作为 CWS 任务的类 ELMO 表示，BERT 特性的作用是什么？

通过对两个基准 CWS 数据集进行广泛的实验，我们得到了几个关于 BERT 的第一手和关键观察，并证明了它在解决 SEC 中的 CWS 任务方面的优势。 3.

2 背景和架构

在下面，我们首先定义了 CWS 任务的问题和令牌化的基本概念。 在此之后，我们提出了 BERT 体系结构并将其应用于 CWS 任务。



(a)变压器 (b)中文分词

无花果。 2. 用于 CWS 任务的变压器和 BERT 体系结构。

2.1 问题定义

中文分词问题定义如下：给定一个输入句子，m 个字符 s=c1c。。c_{*}其中 c 表示第 i 个字符，分段器将用标签 1i 分配每个字符 c，其中“e{B，M，E，S}[17]。 标签 B、M、E 和 S 分别表示单词的开头、中间、结尾和单个字符词。

在本文中，我们采用了在 BERT 的实现中采用的字片标记[16]。 在处理汉字时，字元标记没有任何区别，只有在处理英语单词或数字时略有不同。例如，如图所示。1，the English word，Confluence，is separated into four parts，con，##fl，##ue，and##nce，which yield the corresponding BERT tag of BMME，rather than the source tag of S.For the word，“2004 年”(in2004)，the corresponding BERT tag is BE，where“2004”is annotated by B and“年”(year)is annotated by E.It is fortunate that2004is deemed as a whole word，rather than the case of“2424 位通信院士”(2424communication fellow)，segmented as“2424 位通信院士”，which is tokenized as24，##2，##4，位，通，信，院，士，respectively. 因此，“2424 位通信院士”的源标记由 SSBEBE 表示，其 BERT 标记将分别改为 BMESBEBE。在这个过程中，如果一个字符没有出现在词汇表中，它就被标记为特殊的令牌[UNK]。如果有两个连续的英语单词，我们添加一个特殊的令牌，[未使用 1]，以替换单词中的空格。 这使得我们的过程比以前的方法[1,2,15]更加困难，这些方法将连续数字和英文字符作为一个标记。

如图所示。 另外增加了两个特殊标记[CLS]和[SEP]，分别表示每个句子的开头和结尾，并产生相应的输出标记[START]和[END]。 这两个输出令牌是条件随机场(CRF)的必要令牌，需要对标签之间的依赖关系进行建模。

2.2 伯特

BERT 的基本架构是多层双向变压器编码器，通过调整所有层的左右上下文来学习表示(见图。 2(a)说明，或[14]中的细节)。 最初的预先训练的表示是通过书籍语料库(800M 词)[21]和英语维基百科(2500M 词)上的蒙面语言模型来训练的，而多语言模型则是在 XNLI 数据集上训练的，总共有 112，500 对 15 种语言[5]的注释对。

在序列标记任务方面，给定 m 个字符序列 s=C1C2.CM，我们可以制定 BERT 的体系结构如下：

$$h? \quad =WCi+WP, \tag{1}$$

用=transformer_block(h：1)，我=1，...，L， (2)

=分类器($W_o h^l + b_o$)， (3)

其中 Q 是第一个令牌，W_e 是嵌入层的权重，W_p 是位置编码。 在这里，我们还添加了特殊的令牌，[开始]，作为 c？ 和[END]作为 c_e+1. 是层数 transformer_block 层，由 sel^注意和完全连接层[14]W 组成. 。和 b_e 分别是输出层的权重矩阵和偏置。分类器可以是 CRF 或 Softmax。

在我们采用的 BERTbase 中，L=12。 我们是 R|其中 H=768 和|D 通过在中文集合上应用 Bertbase 模型=21128，该模型由 21128 个英文、中文、表情符号和一些特殊符号组成。 位置编码 $WpeR^{n\times1}$ 最大序列长度为 512。 输出权重矩阵 $W_e R^{r\times H}$ 和 $b_e R^{r\times1}$ 其中 T 是输出标签的数量，即我们测试中的 6。

表 1。 数据集的统计。 ‘#’符号代表“的数目‘。

数据集	部分	#发送。	#单词	#Chi。言语	#Eng。言语	#数字	#Chars	奥夫
Msr	火	87k	2368k	2350k	1,154	18K	4050k	1,991
	测	4k	107k	106k	66	697	184k	0.023
pku	火	19k	1,110k	1090K	443	20k	1,826	2,863
	测	2k	104k	102k	28	2k	173k	0.05

表 2。 与以前的模型比较

方法	Msr				pku			
	CRF		Softmax		CRF		Softmax	
	f1	A	f	Acc。	f1	Acc。	f1	Acc。
SE+半 CRF[9]	97	—	–	–	**96.**	–	--	
WCC 嵌入 +CRF[20]	97 .8	—	–	–	96. 0	–	–	–
我们 +CNN+CRF[15]	98 .0	—	–	–	96. 5	–	–	–
CE+Bi LSTM +Softmax[10]	98 .1				96. 1			
伯特	**98**	**9**	**98**	**99.0**	96.	**97.6**	**96.**	**97.6**

3 实验

数据。 我们在第二个国际中文分词烘焙[7]的 PKU 和 MSR 两个基准数据集上评估 BERT。 数据集的统计数据如表 1 所示。

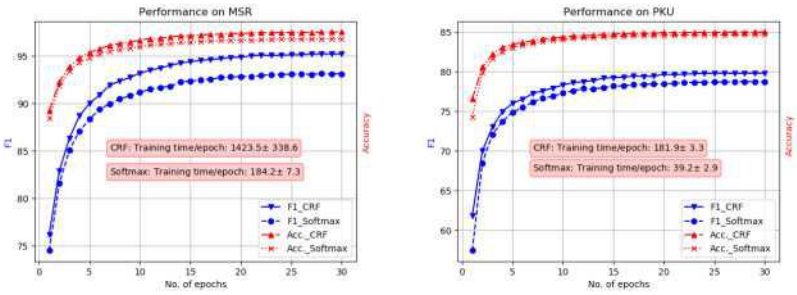
评价。 标准词 F1 测度[7]用于评价分割性能。 此外，我们还计算了精度，以评估更多方面的性能。

设置和设置。实验在 40 核 IntelXeon CPUE5-2630v4@2.20GHz 和 128G 内存的服务器上运行，模型在 NVIDIATITANXP 图形卡的 12G 内存上训练，共由四个 GPU 组成。

采用中文语料库训练的 BERTbase 作为输入模型，由 12 个 BERT 层组成，隐藏大小为 768，自注意头数为 12，参数约为 110M。在对模型进行微调时，我们采用 ADAM[8]作为优化器。 学习率设置为 2e-5。 最大序列长度设置为 128。

3.1 主要成果

表 2 列出了最近应用的基于神经网络的模型的最新结果，以及 BERTbase 功能调谐的性能。 观察到 CRF 和 Softmax 具有相同的性能，在 MSR 数据集中，两个分类器在所有模型中都达到了最佳性能，而在 PK U 数据集中，Softmax 达到了最佳性能，CRF 在比较模型中达到了竞争性能。 在 Softmax 分类器方面，细化 BERTbase 可以分别在 MSR 和 PKU 数据集上进一步提高+0.3F1 分数和 0.4F1 分数。



(a) MSR (b) PKU

无花果。 3. 仅在第一/嵌入层上进行 BERT 微调的消融研究。

3.2 分类器的效果

我们研究了分类器在这两个数据集中的效果。 在图中。 我们只从第一层，即嵌入层报告了对特征进行微调的结果。 结果表明

这两种分类器的性能都随着时代的增加而逐渐增加和收敛。 总的来说，当只在一层进行微调时，CRF 的性能优于 Softmax。

通过观察表 3 中的结果，我们可以发现 CRF 和 Softmax 之间的差距变小，当层的

大小为 12 时，它们达到相同的水平。 这些结果表明，当提取的特征不够时，更复杂的分类器(CRF)可能有助于性能。

我们还注意到，CRF 的每个时期的时间成本远大于 Softmax。 在 MSR 数据集中约为 7.7 次（1423.5 秒对 184.2 秒），在 PK U 数据集中约为 4.6 次(181.9 秒对 39.2 秒。 当 BERT 模型尺寸变大时，我们可以提取更多的特征。 结果表明，CRF 和 Softmax 之间的性能差距变得微不足道。 因此，我们建议 Softmax 作为最终分类器，因为它的简单性。

3.3 模型大小的影响

我们探讨了模型大小对不同层数的 BERT 模型的功能调整的影响。 由于内存不足的问题，当层数分别为 1、3、6 和 12 时，我们将批处理大小设置为 384、128、64 和 32。

从表 3 中的结果，我们观察到这一点

当模型大小（层数）增加时，性能逐渐增加。

当层数较小时，CRF 通常比 Softmax 获得更好的性能，除了 PKU 数据集中 L=6 时的情况。 我们推测这是因为 PKU 数据集是一个相对较小的数据集

表 3。 在 MSR 和 PKU 数据集上对不同层数(以#L 表示)的 BERT 进行了消融研究。

#l	Msr		pku	
	CRF		CRF	
	F1Acc。	F1Acc。	F1Acc。	F1Acc。
1	95.2 97.5	93.1 96.8	79.8 85.0	78.8 84.7
3	97.5 98.6	97.3 98.6	84.6 88.2	83.8 87.3
6	98.2 98.9	98.1 98.9	85.6 88.5	86.9 89.9
1	98.4 99.0	98.4 99.0	96.5 97.6	96.5 97.6

具有较大的词汇量外(OOV)率，这使得模型没有得到很好的训练。

表 4。 在 MSR 和 PKU 数据集上采用基于特征的方法对 BERT 进行了消融研究。 从指定层的激活被组合并输入到两层 BiLSTM 中，而不更新 BERT 中的权重。

分层	Msr				pku			
	CRF		Softmax		CRF		Softmax	
	f1	Ac	f	Acc。	f1	Ac	f1	Acc。
全部完成	98	99	9	99.0	96.	97	96	97.6
第一层（嵌入）	95	97	9	97.1	91.	94	91	94.2
二到最后隐藏	96	97	9	97.3	94.	96	94	95.8
最后的隐藏	96	97	9	97.1	82.	85	93	95.5
最后四个隐藏	96	98	9	98.0	95.	96	94	96.4
简洁的最后四个隐藏	96	98	9	98.0	95.	96	94	96.4
共 12 层	97	98	9	98.2	95.	96	95	96.6

3.4 基于特征的结果

我们还通过在 CWS 任务上生成类似 ELMO 的[12]预先训练的上下文表示来评估 BERT 在基于特征的方法中的表现。 为了做到这一点，我们应用了一个或多个层的激活，而不需要对 BERT 的任何参数进行函数化。 这些上下文嵌入被用作在分类层之前随机初始化的两层 BiLSTM 的输入。 从表 4 中报告的结果，我们观察到这一点

最佳性能是通过对预先训练的变压器的所有 12 个隐藏层的表示进行求和来实现的，它在 F1 后面是 1.3 和 1.5，这是通过对 MSR 数据集中的所有 12 个隐藏层进行函数调整而得到的，而 PKU 数据集中的间隙是 1.3 和 1.4。结果还证明了 BERT 在基于特征的方法中的优势。

使用第二到最后隐藏层的性能通常比使用最后一个隐藏层要好。 这意味着最后隐藏层中的激活不适合最终的下游任务。

当 CRF 应用于 PK U 数据集中最后一个隐藏层的激活时，性能最差。 这再次表明，当训练集小时，CRF 是不合适的。

3.5 消融研究

为了了解 BERT 得到的结果，我们还从 MSR 和 PKU 测试集中随机选择一些错误，并手动分析它们。

与[10]中的观察相似，在 MSR 测试集中，BERT 将抽象概念（抽象概念)分别视为抽象(抽象)概念(概念)，因为象(抽象)在 MSR 训练集中作为一个词出现了 30 次。与[10]中的观察不同，在 MSR 测试集中的相关权（权利/权力）一词中，BERT 只对统治权（统治权力)的情况犯了错误，并分别将其作为统治(统治权力）和权(权力。对于其他情况，“审批(vetting)权(right)”，“建筑(construction)权(right)”，“领导(leader)权（功率）”，BERT 正确地将单词分割为标记数据，这表明 BERT 在分割单词“统治权（统治权力)”时的一致性）”。 在 PKU 测试集中，BERT 将在标记测试集中分别将“关税权（关税权)”、“贸易权(贸易权)”、“航行权(航向权)”、“诉权(仅要求)”、“关税(关税)”和“权(权利)”、“贸易(贸易)”和“权(权利)”、“航行(航向)”和“权(权利)”、“诉(公正）”和“权(索赔。 结果表明，BERT 在训练集中一致地将单词分割为相同的标准，而不是测试集中的不一致。在 PKU 测试集中，BERT 对“有权有势（具有权力和影响力)”和“位咼权重(至高无上和强大)”这两个词进行分段，这比手动标记的单词要好得多：“有权(具有权力)”、“有(拥有)”、“势(影响力)”；以及“位(位置)”、“高(高)”、“权(权力)”、“重(重量）”。

在[10]中与“县（县）”相关的单词方面，在 PKU 测试集中，BERT 在分割相关单词方面没有显着性差异。 只有两种情况是将一个词分成两个词，例如，“堆龙德庆县（德清县)”被“堆龙(德清县)”和“德庆县(德清县)”，“县区(县)”被“县(县)”和“区(区）”。 同时，三种情况是将两个词合并为一个词，例如，“市县（市和县)”和“级(级)”合并为“市县级(市和县)”；“县(县)”和“政府(政府)”合并为“县政(县政府)”；“先进(先进)”和“县(县)”合并为“先进县(先进县）”。 我们觉得这个组合使单词更加紧凑。

在 MSR 测试集中，“县（县)”将与“穷(贫困)县(县)”、“县(县)(中国人民银行)”、“县(县)消委会(消费者委员会)”分开）”。 在 BERT 预测中，它们对应于“穷县（贫困县)”、“县人行(中国人民银行）”，

和“县消委会（县消费者委员会）”，分别。 BERT 在分割这句话时只犯了一个重大错误，“本报发表了记者在山东茌平县采写的调查报告(该报纸发表了一份由记者在山东齐平县撰写的调查报告。)”。

—The sentence is manually segmented as“本报(the newspaper)/发表(publish)/了(ed，past tense)/记者(reporters)/在(in)/山东(Shandong Province)/茌平县(Chiping County)/采写(written)/的(of)/调查(survey)/报告(report)”。

—However，the BERT result is“本报(the newspaper)/发表(publish)/了(ed，past tense)/记者(reporters)/在(in)/山东(Shandong Province)/[UNK]平县([UNK]ping County)/采写(written)/的(of)/调查(survey)/扌艮告(report)”。

虽然这个例子没有受到影响，但未知的令牌(UNK)通常会使 BERT 误解整个句子，需要额外的后处理。

表 5。 通过 BERT 分别预测 MSR 测试集中的 Idoms。

黄金	伯特
神经衰弱神经衰弱	神经 nerve/衰弱弱
若有所思 thoughtful	若如果/有所有某事。 /思想
重男轻女 patriarchal	重宝藏/男男孩/轻处置/女女孩
崖崖畔畔 cliffside	崖悬崖/崖悬崖/畔侧/畔侧
男女平等两性平等	男女男女/平等平等
人定胜天人将赢得这一天	人人/定确保/胜赢天/天
另眼相看方面给予特别注意	另另一个/眼的眼睛/相看盯着对方
不可偏废不可忽略	不可不能/偏部分/废放弃
一好百好对所有人都有好处	一好百好一百/好好
夕卜弓丨内联引进国外投资，在国内建立横向联系	夕卜外部/弓丨介绍/内内部/联团结
收获时以丰补歉储存，不足时弥补	以拿/丰富/补补/歉穷

通过探讨分割结果中的其他差异，我们观察到它们在于分割 idoms。 我们将它们分别列在表 5-8 中，并进行如下观察和猜想：

表 6 中的所有单词都没有出现在 MSR 训练集中。 我们推测它们来自在 BERT 中训练的外域资源。

-在表 8 中，th 乞 words、“银装素裹”、“不懈努力”、“假冒伪劣”、“至关重要”、“难以为继”、“受益匪浅”、“”和“证据确凿”在 PKU 数据集中的训练集和测试集中的标记不一致。 显然，BERT 适合于训练集，并在测试集中进行不同的预测。

换句话说，“徘徊不前”、“天真无邪”、“倾囊相助”、“喜中有忧”和“心知肚明”不出现在训练集中，可能来自在 BERT 中训练的外域资源。

表 6。 MSR 测试集中的 Idoms 由 BERT 作为一个整体预测，但单独标记。

伯特	黄金
新春佳节春节	新春新的春天/佳节节日
胜券在握胜利在握	胜券对胜利/在握持有的信心
整装待命准备待命	整装准备好东西/待命等待命令
东升西落东起，西落	东东/升升/西西/落套
圆缺盈亏 wax 和万	圆圈/缺缺少/盈亏蜡和 wan
乱采滥伐森林砍伐	舌 L 无序/采矿井/滥伐毁林
稳中求进寻求稳定性的改善	稳稳定/中中/求寻求/进改进
心有余力有多余的能量	心心/有余有剩余/力力量

表 7。 通过 BERT 分别对 PKU 测试集中的 Idoms 进行预测。

黄金	伯特
洛洛大方自然优美	落落落/大方大方
长夜漫漫 long 晚	长夜长夜/漫漫很长
取信于民赢得人民的信任	取信于赢得信任/民的人
杀人火口谋杀某人。 防止泄露自己的秘密	杀人谋杀/火口带走了证人
刑讯逼供用酷刑强迫一份声明	开 U 讯刑讯逼供/逼供逼供
匡扶正义伸张正义	匡扶维护/正义正义

4 结论

在本文中，我们进行了广泛的实验，以研究 BERT 在解决 CWS 任务中的作用。 从结果中发现了几个 oberservations：

紧急状态可以稍微提高 CWS 任务的性能。 更具体地说，在 Softmax 实现的 F1 评分方面，MSR 数据集和 PKU 数据集分别+0.3 和 0.4 增益。

当应用充分学习的特征时，CRF 和 Softmax 达到相同的性能。 然而，由于时间成本低，Softmax 更受欢迎。

随着模型尺寸的增大，BERT 的性能逐渐增大。

—通过 BERT 提取的特征也可以作为其他神经网络模型的良好候选。

在 PK U 数据集中对 idoms 的分析结果帮助我们在数据集中找到标记的不一致问题。

对于这样的预测错误，是无法纠正的。

有几个有前途的研究方向与我们的工作有关。

首先，我们目前的实现不能很好地处理由汉字和英语单词组成的多语句。 设计新的机制来处理它们是实用的。

—第二，OOV 是一个关键问题，因为它会产生一个未知的令牌，这混淆了 BERT 对汉语单词的分割。 似乎用更多的中国资源训练一个新的 BERT 模型是一个潜在的解决方案。

第三，目前的工作旨在解决域内 CWS。 有希望探索有效的方法，使受过训练的模型适应新的领域，例如社交媒体，它由简短的文本和特殊的标记组成。

表 8。 在 PKU 测试集中，Idoms 被 BERT 作为一个整体预测，但单独标记。

伯特	黄金
银装素裹穿着银白色的衣服	银装银/素平/裹包装
不懈努力不合理的努力	不懈不合宜/努力的努力
假冒伪劣伪造和假冒商品	假冒假/伪劣假
非常重要的至关重要	全夫相当/重要的重要性
难以为继不可持续	接班难以困难/为继
受益匪浅 benefit 很多	受益好处/不匪/浅浅薄
徘徊不前 not 蹲着	徘徊悬停/不不/前面囙 U
天真无邪天真纯洁	天真无辜/无没有/邪邪恶
蔚为壮观 spectacular	蔚为负担得起/壮观壮观
倾囊相助尽全力帮助某人。	倾倒出/囊包/相助帮忙
喜中有忧没有欢乐就没有烦恼	喜开心/中中/有有/忧担心
一动一静一动一静	一一/动移动/一一/静安静
闻风而止 smell 风	闻闻/风风/而而且/止停止
心知肚明某事。 我已经知道了	心心/知知道/肚胃/明明白
证据确凿无可辩驳的证据	证据证据/确凿无可辩驳

承认

本文所述工作得到中国香港特别行政区研究资助局（项目编号）的部分支持。 UGC/ids14/16)。

参考资料

1. 蔡，D，赵，H：汉语神经分词学习。 In：ACL（2016）
2. 蔡，D，赵，H，张，Z，辛，Y，吴，Y，黄，F：汉语快速、准确的神经分词。 In：ACL。 pp。 608-615（2017）
3. 陈，X，邱，X，朱，C，黄，X：门控递归神经网络用于汉语分词。In：ACL。pp。1744-1753（2015）
4. 陈，X，邱，X，朱，C，刘，P，黄，X：用于汉语分词的长期短期记忆神经网络。 In：

EMNLP。 pp。 1197-1206（2015）
5. Conneau，A.，Rinott，R.，Lample，G.，Williams，A.，Bowman，S.R.，Schwenk，H.，Stoyanov，V.：XNLI：评估跨语言句子表示。 In：EMNLP。 pp。 2475-2485（2018）
6. Devlin，J.，Chang，M.，Lee，K.，Toutanova，K.：BERT：深度双向变压器的预训练，用于语言理解。 In：NAACL-HLT（2019 年）
7. 艾默生：第二届国际汉语分词贝科夫。 In：SIGHAN@IJ CNLP（2005）
8. 金马，D.P.，Ba，J.：亚当：一种随机优化方法。 In：ICLR（2015 年）
9. 刘，Y，车，W，郭，J，秦，B，刘，T：探索神经分割模型的分段表示。 In：IJ CAI。 pp。 2880-2886（2016）
10. 马，J，甘切夫，K，魏斯，D：最先进的中文分词与双 LSTMS。In：EMNLP。pp。4902-4908（2018）
11. 裴，W，葛，T，张，B：中国分词的最大裕度张量神经网络。 In：ACL。 pp。 293-303（2014）
12. 彼得斯，M.E.，Neumann，M.，Iyyer，M.，Gardner，M.，Clark，C.，Lee，K.，Zettlemoyer，L.：深度语境化的词表示。 In：NAACL-HLT。 pp。 22272237（2018）
13. Radford，A.，Narasimhan，K.，Salimans，T.，Sutskever，I.：通过生成性预训练来提高语言理解。 技术。 代表，开放人工智能（2018 年）
14. Vaswani，A.，Shazeer，N.，Parmar，N.，Uszkoreit，J.，Jones，L.，Gomez，A.N.，Kaiser，
L. 波洛苏欣：你只需要注意。 In：NIPS。 pp。 6000-6010（2017）
15. 王，C，徐，B：具有词嵌入的卷积神经网络用于汉语分词。In：IJ CNLP。pp。163-172（2017）
16. Wu，Y.，Schuster，M，Chen，Z.，Le，Q.V.，Norouzi，M.，Macherey，W.，Krikun，m. 曹，Y，高，Q.，Macherey，K.，Klingner，J.，Shah，A.，Johnson，M.，Liu，X.，Kaiser，L.，Gouws，S.，Kato，Y.，Kudo，T.，Kazawa，H.，Stevens，K.，Kurian，G.，Patil，N.，Wang，W.，Young，C.，Smith，J.，Riesa，J.，Rudnick，A.，Vinyals，O.，Corrado，G.，Hughes，M.，Dean，J.：Google 的神经机器翻译系统：人与机器翻译之间的差距。 共同 RRabs/1609.08144（2016）
17. 徐，N：中文分词作为字符标注。 ijclclp8（1)(2003 年）
18. 杨，J，张，Y，董，F：神经分词与丰富的前训练。 In：ACL。 pp。 839-849（2017）

19. 赵，H，蔡，D，黄，C，Kit，C：中文分词：另一个十年回顾（2007-2017）。 公司 RRabs/1901.06079（2019）
20. 周海，余，Z，张，Y，黄，S，戴，X，陈，J：中文分词的构词特征嵌入。 In：EMNLP。 pp。 760-766（2017）
21. Zhu，Y.，Kiros，R.，Zemel，R.S.，Salakhutdinov，R.，Urtasun，R.，Torralba，A.，Fidler，S.：使书籍和电影对齐：通过看电影和阅读书籍来实现类似故事的视觉解释。 In：ICCV。 pp。 19-27（2015）