基于双编码器的医学文本中文分词

宗源[1,2]，常宝宝[1,3*]
（1. 北京大学计算语言学教育部重点实验室，北京100871；
2. 北京大学软件与微电子学院，北京102600；
3. 鹏城实验室，深圳518055）
{zongyuan,chbb}@pku.edu.cn

摘要

中文分词是自然语言处理领域的基础工作，然而前人的医学文本分词工作都只是直接套用通用分词的方法，而医学文本多专用术语的特点让分词系统需要对医学专用术语和医学文本中的非医学术语文本提供不同的分词粒度。本文提出了双编码器医学文本中文分词模型，利用辅助编码器为医学专有术语提供粗粒度表示。模型将需要粗粒度分词的医学专用术语和需要通用分词粒度的文本分开，在提升医学专用术语的分词能力的同时最大限度地避免了其粗粒度对于医学文本中通用文本分词的干扰。

关键词： 医学文本信息处理 ； 中文分词

Chinese word segmentation of medical text based on dual-encoder

ZONG Yuan[1,2], CHANG Baobao[1,3*]
（1. Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing 100871, China;
2. School of Software and Microelectronics, Peking University, Beijing 102600, China;
3. Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China）
{zongyuan,chbb}@pku.edu.cn

Abstract

Chinese word segmentation is the basic work in the field of natural language processing. However, the previous medical text segmentation work only directly applies the general word segmentation method, and the characteristics of medical texts containing medical terminology make the word segmentation system need to deal with medical terminology and medical text. Non-medical term text provides different segmentation granularity. This paper proposes a dual-encoder medical text Chinese word segmentation model, using an auxiliary encoder to provide a coarse-grained representation of medical-specific terms. The model separates medical-specific terms that require coarse-grained word segmentation from texts that require general-purpose word segmentation, which improves the word segmentation ability of medical-specific terms while avoiding the interference of its coarse-grained word segmentation in general texts in medical texts.

Keywords: Medical text information processing , Chinese word segmentation

*通讯作者：chbb@pku.edu.cn

第二十届中国计算语言学大会论文集，第76页-第85页，呼和浩特，中国，2021年8月13日至15日。

1 引言

当前，针对通用文本的中文自动分词技术已经较为成熟。无论是基于机器学习的传统方法还是深度学习兴起之后的LSTM-CRF模型，再到以预训练模型为基础实现的中文分词系统在通用领域都已经达到了不错的效果。但是，针对医学领域的中文分词却并没有类似完善的处理方式。随着大数据技术的发展，智慧医疗逐渐进入了我们的生活，由于中文分词对智慧医疗中许多后续任务的帮助极大，对于医学文本进行中文分词的需求也越来越迫切。

本文的中文分词任务基于面向医学文本处理的医学实体标注规范定义 (Zhang et al., 2020)，如表1所示。为了更好地在后续任务进一步使用分词处理结果，将医学文本中的除去症状之外的医学命名实体看作一个整体。这些命名实体可以是医学领域专有的疾病名、检查程序或是身体物质，在医学文本中有大量这样的医学专用术语。医学专用术语的理解和切分对人类来说都很困难，对模型来说更是如此。而这类模型难以理解的医学专用术语，人们往往关注的是其整体的含义，而且往往在后续任务中需要将其作为一个整体进行分析。因此为了便于后续任务的处理，在进行中文分词时研究者对大多数的医学命名实体在内部不予以切分。

文献/报道/161/例/**原发性肺泡换气不足综合征**/患儿/有/27%/伴发/HD/。 9/例/**直肠活体组织检查**/标本/为/**神经节细胞症**/。 **全结肠无神经节细胞**/患儿/结肠/可/无/典型/狭窄/表现/。

Table 1: 医学文本中文分词示例

表1例中的“原发性肺泡换气不足综合征”、“直肠活体组织检查”、“神经节细胞症”、“全结肠无神经节细胞”均为医学专用术语，这些术语所需的分词粒度同通用文本明显不同。在医学文本中医学术语的分词需要粗粒度的同时，数据集文本中的其他词汇却需要使用常规分词粒度进行分词，这里就出现了分词粒度区分的需求。我们希望模型能够分辨文本中的医学专有词汇并将其作为一个整体看待，让模型在医学专有词汇内部不予以分词，而在通用的命名实体内部正常进行切分。然而对于通用文本分词系统来说，如何在医学文本中自动发现医学命名实体，并对其使用粗粒度分词就成了一个棘手的问题。

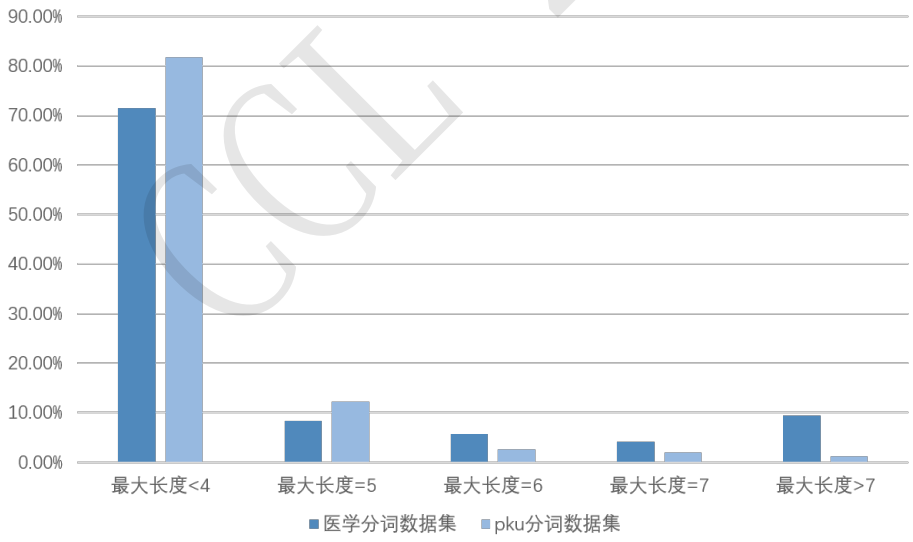

Figure 1: 单句最长词汇长度统计

我们发现，只有长度较长的医学专用术语会需要粗粒度分词，这样的术语可能包含着大量的前缀或是后缀信息，而从传统分词的角度来看，前后缀和词语主体是可以进行切分的。对于长度较短的医学专有术语来说，将它们看作一个整体的分词粒度同通用文本的分词粒度并没有

差异。我们统计了根据张欢 (2020)等的医学命名实体标注规范标注的医学文本分词数据集和通用分词数据集的每句话的最长词汇的长度分布，如图1。统计发现医学文本数据集的词汇长度分布同通用分词数据集不同，由于包含医学专有术语，医学文本数据集中每句话的最长词语长度较通用文本数据集更长。其中医学文本分词数据集包含长度大于等于7的词语的句子在数据集中占比约为10%，这些词语绝大多数都是医学专用术语。在数据集中，这类医学专用术语和通用词语混在一起，如果只是简单使用通用分词方法的单一分词粒度进行分词，训练集中医学专用术语和通用词语的不同分词粒度将会互相带来干扰。这种干扰不仅会在医学专用术语内部出现不必要的切分，训练集中粗粒度的专用术语也会影响到模型对通用词语的切分，导致整体模型效果的下降。

针对通用中文分词系统无法针对医学专用术语提供专用粗粒度表示的特点，本文提出了双编码器的中文分词系统。与通用领域的分词方法相比，模型添加了辅助编码器以提供粗粒度分词表示，同时使用判别器判断输入文本是否含有医学专用术语。当输入文本包含医学专用术语时，辅助编码器将参与分词过程，提供粗粒度的分词表示，再将粗粒度的分词表示和主编码器提供的编码表示结合进行序列标注。由于辅助分类器只在输入文本中包含医学专用术语时参与梯度回传，使得其能够产生包含医学术语的粗粒度分词表示。实际使用时，模型首先判断输入文本中是否包含医学术语，再根据判断结果决定是否将两种表示进行组合，最终可以让模型为不同类型的输入提供不同粒度的中文分词。实验结果表明，在现有的医学文本中文分词数据集上我们提出的双编码器模型在各项指标上均超过了基线模型。

2 相关工作

中文分词属于序列标注任务，是中文信息处理的一个基础方向，也是进行其他中文自然语言任务的基础。由于中文文本中词与词之间并没有明确的分割标记，而是直接相连成为一个连续字符串。在自然语言处理领域，分词作为后续任务的基础，直接影响着后续任务的效果。现有的医学分词系统均基于通用分词方法实现。张立邦 (Zhang et al., 2014)提出了一种无监督分词方法解决中文电子病历的分词，模型首先利用通用词典对文本进行初步切分，再利用字串的左右分支信息熵使用EM算法构建良度对初步的切分结果进行调整，以达到识别未登录词的目的。王若佳 (2019)使用现有的开源分词工具和第三方资源进行了中文病历的分词和命名实体，其中分词模型使用Jieba、Jieba+用户词典、无监督学习和AC自动机实现了对病历文本的分词。王莉军 (2020)使用Bi-LSTM和Bi-LSTM-CRF两种深度学习模型实现了中医数据的中文分词，并把这两种模型和现有基于统计方法的开源分词工具Ansj和Jieba进行对比，得到了优于现有开源分词工具方法的效果，证明了深度学习在医学文本分词任务上的有效性。然而，现有的医学文本分词系统也只是将通用的分词方法直接在医学文本上使用，并没有考虑医学文本与通用文本之间的差异。

可以发现，现有的医学文本中文分词系统大多是基于通用文本的中文分词，这些通用文本分词方法主要分为基于词典和统计的方法和基于深度学习的中文分词方法。

基于词典匹配的中文分词，顾名思义就是将需要分词的语段用一个足够大的词典进行匹配。在针对语段的每个范围中，如果在词典中找到了对应的某个词汇，则成功配对，最终找到一个最满足的分词结果。随着统计方法在自然语言领域的应用，研究者逐渐开始使用统计方法解决中文分词任务，Xue (2003)提出将分词任务看作一种序列标注任务，把字在词中的相对位置作为字的标签进行训练，即B、M、E和S的标签分别代表词的开始、中间、结束和单个字的词，从而让分词任务成为了序列标注任务并被沿用至今。之后 (Ng and Low, 2004)第一次使用BMES的标注方式使用最大熵算法进行了中文分词。而后，由于最大熵马尔科夫在每个节点单独归一化会导致标记偏置问题，自然地，Peng (2004)和Tseng (2005)将条件随机场CRF (Lafferty et al., 2001)引入来解决中文分词问题。随后，由于CRF的优秀表现，由CRF的多种变种成为了深度学习之前解决中文分词的标准方法。

基于深度学习的中文分词也被称作基于理解的分词方法，这种方法希望计算机模拟人对文本的理解之后对文本进行分词，它模拟了人对句子的理解过程。2013年，Mairgup (2013)提出使用特征结合神经网络的方法解决中文分词问题，首次验证了在中文分词任务上使用深度学习方法的可行性。同年，Zheng (2013)也在分词任务中使用深度学习方法，使用大规模预训练的字向量作为输入，并使用类似感知机的训练方式加速神经网络训练。Pei (2014)在2014年引入了标签向量对Zheng的模型做出改进，并在文中提出了一种新型的张量分解方式。Chen (2015)提

出了使用自适应门结构的递归神经网络（GRNN）来提取n-gram特征。同年，Chen (2015)使用长短时记忆神经网络（LSTM） (Hochreiter et al., 1997)来捕捉长距离的依赖，部分解决了之前只能从滑动窗口提取特征的不足。Xu和Sun (2016)将GRNN和LSTM联合起来训练模型，模型先用双向LSTM提取上下文信息，再使用滑动窗口将这些提取出来的信息使用带门结构的递归神经网络融合起来，最后进行标签的分类工作。之后随着深度学习在自然语言处理方向的发展，Ma (2018)发现只需要使用Bi-LSTM模型，添加各种的预训练模型、dropout和参数调优就可以将分词提升到领先水平。实验发现主要的性能瓶颈来自于未登录词的识别，于是在模型中添加更多的外部信息成为了分词的发展趋势。Zhang (2018)在表示中融入词典外部信息得到了提升的效果。Wang (2018)在语义表示的基础上，添加了字的拼音和五笔等特征，使用Bi-LSTM-CRF模型训练证明多特征融合对分词效果的提升。未来使用更好的预训练模型和更有效的特征融合都将是中文分词的重要研究方向。

综上，当今中文分词研究都是基于通用分词任务上进行的，这样训练出的模型虽然在各大通用数据集上测试表现出色，但是直接将这些方法运用在医学文本分词任务上由于没有考虑到医学文本的特殊性并不能达到与在通用文本中类似的效果。现有的医学文本中文分词系统也只是将通用分词算法直接用于医学文本分词。但是医学文本中不仅包含通用文本，还包含大量与通用文本所需分词粒度不同的医学专用术语，这需要分词系统根据输入自动提供不同的分词粒度。但是当前的通用分词系统并没有这种动态辨别文本中是否包含医学实体的能力，于是模型在对于包含医学命名实体的分词时就会带来分词粒度上的问题。这个问题导致通用分词方法中对所有文本进行一次编码的方式不适合对医学文本进行分词，无论使用粗粒度分词还是细粒度的分词都会带来分词错误。因此，本文针对医学文本中医学术语需要的特有分词粒度提出了双编码器分词模型。模型改变了通用分词系统对于输入文本只有一种表示的情况，可以自动发现输入文本中医学术语的存在并对医学术语的表示进行优化。

3 模型结构

本文的提出的双编码器医学中文分词模型由两个编码器和序列标注模块组成，其中一个编码器是主编码器，负责生成通用文本分词粒度的编码表示，另一个则负责表示输入文本中可能带有的医学术语信息，当输入文本中含有医学术语时将辅助编码器输出的表示添加到主编码器的表示中。

为了判断输入中是否包含医学专用术语，我们设计的判别器模块将从两个编码器分别获得整个输入的两种表示，并将它们合并得到输入的信息表示。判别器使用这个最终的信息表示对输入文本进行分类，判断这段输入文本中是否包含医学术语。再根据结果，如果输入文本中包含医学术语则将医学术语的表示加到常规表示上，从而提升了只使用单编码器对于带有医学术语的文本的表示能力。最后将经过双编码特征提取器处理后的结果经过一个双向LSTM网络和CRF层对每个字符输出序列标注预测。

输入包含医学文本$X = [x_1, x_2, \ldots, x_n]$，以及是否包含医学术语的标记$c = \{0, 1\}$。模型输出即为输入文本的标签$Y = [y_1, y_2, \ldots, y_n]$，其中$y_i = \{B, M, E, S\}$。模型具体由三个部分组成，分别是双编码特征提取器，医学术语判别器和序列标注模块，下面分别阐述这三个部分。

3.1 双编码特征提取器

本文使用两个编码器模块处理输入文本，其中一个是主要的表示编码器S，另一个是辅助的医学术语表示编码器M，模型如图2。

编码器负责自动提取输入文本的特征，每个字符作为一个输入单元输入编码器模块中后得到各自的表示向量，输出为包含字符内容信息的表示向量。我们将这个输出当作每个输入字符的特征表示，并将作为后续序列标注模块的输入。本文通过辅助的医学术语编码器提取语句中可能包含的医学术语特征，于是在标签预测的过程中可以将医学术语特征考虑进来。同时为了后续对于输入文本的整体判断，模型在每段输入文本之前添加一个$[CLS]$标签代表整句话的表示，模型中这个标记所对应的输出表示会在后续被用来判断该句输入文本中是否包含医学术语。

双编码器中的每一个编码器都可以根据输入得到一组表示。其中编码器S得到的是每个输入字符的意义表示$H_s = [h_1^s, h_2^s, \ldots, h_n^s]$，负责生成常规粒度分词的文本表示。而编码器$M$得到的是包含医学术语时的每个输入字符的表示$H_m = [h_1^m, h_2^m, \ldots, h_n^m]$，将生成更倾向于将文本进

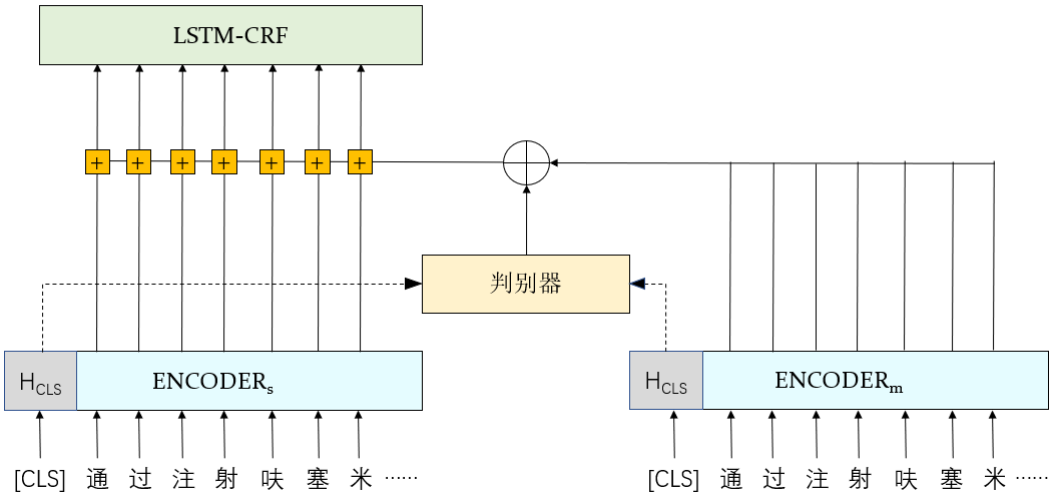

Figure 2: 双编码器中文分词模型

行粗粒度分词的表示。这个编码器只在输入文本中包含医学术语时才参与后续的序列标注任务，由此保证该编码器学习到的是包含医学术语的分词表示。

$$H_S = ENCODER_S(token) \tag{1}$$

$$H_M = ENCODER_M(token) \tag{2}$$

这两种输出将根据医学术语判别器的分类结果决定如何在序列标注模块中使用。

3.2 医学术语判别器

由于包含医学术语的医学文本分词包含长度较长的词语，本文认为医学专用术语所需的分词粒度将大于数据集中的通用文本。但是，为了判定是否需要辅助编码器为输入提供粗粒度分词表示，需要判断文本中是否含有医学专用术语。医学术语判别器由一个两层的前馈神经网络组成，负责判断输入文本中是否包含医学专用术语。我们将两个编码器在$[CLS]$标签所在位置的表示取出，作为每段输入文本在通用分词情况下和需要粗粒度分词情况下的两种整体表示，分别是$[CLS]_s$和$[CLS]_m$，再将这两个表示合并后经过MLP分类器得到文本中是否包含医学术语词汇标签C，其中$C = \{0, 1\}$。

$$C = Concat([CLS_s, CLS_m]) \tag{3}$$

3.3 序列标注模块

序列标注模块的输入包含双编码特征表示器输出的h_s和h_m两种表示以及医学术语判别器的输出C。模型的序列标注模块由表示组合器根据医学术语判别器的结果后接LSTM-CRF模块组成。

3.3.1 表示组合器

由于医学术语和通用文本的分词粒度不同，因此包含医学术语的文本表示应当在通用表示中加入医学术语分词所需的粗粒度表示。当输入文本中包含医学术语时，我们将编码器M的输出同编码器S的输出相加形成包含医学术语表示的文本编码表示。当输入文本不包含医学专有术语时，则直接使用编码器S的输出作为输入文本的编码表示。根据医学术语判别器的输出C，我们将医学文本的表示组合成：

$$Representation_{final} = \begin{cases} H_s & C = 0 \\ H_s + \lambda H_m & C = 1 \end{cases} \tag{4}$$

3.3.2 LSTM-CRF序列标注模块

在以BERT为代表的预训练模型兴起之后，使用预训练模型作为基础编码器解决序列标注任务已经成为主流。由于现有的特征编码器都基于Transformer (2017)结构，而Transformer的自注意力（Self-attention）机制仅仅利用位置嵌入（position-embedding）将每个字符的位置信息传入模型，因此在编码器之后添加传统的LSTM学习输入序列的依赖关系依然十分必要。在预训练编码器后添加LSTM层,可以强化模型的位置信息，以进一步完善输入字符的特征表示。另外，如果只用编码器表示经过LSTM输出的结果表示来预测每个字符的标签则仅仅考虑了每个字符输入本身的信息，却无法学习到序列每个输出标签之间的关系。在序列标注任务中，每个字符的标注之间是有一定关系的，例如，标签B后不会再出现一个标签B或是S，而一般会接标签E或是M。在模型最后加入CRF层则很好地解决了这一问题，CRF是在全局范围统计归一化的条件状态转移概率矩阵，相当于对编码信息的再利用，考虑输入表示和标签关系为整个文本输入预测出每个输入的最优标签。综上，本文的序列标注模块选择在模型中添加两层的LSTM网络，最后使用CRF层经过解码输出预测结果。

4 数据来源

本文使用的医学文本分词数据集是根据张欢 (2020)等的医学命名实体标注规范标注的医学文本分词数据集，我们根据数据集中包含的句子数量进行了数据集的切分，在按照9:1:1的句子数量比例将数据集分割为训练集、开发集和测试集。

另外，我们根据分词数据的标注结果，给每一个输入样例标注是否包含需要粗粒度分词的标签，我们发现数据集中词语长度大于等于7的词为绝大多数都是医学术语，于是我们将包含这样的词的样例标注为包含长医学术语，并将会在模型中为这类样例提供粗粒度分词表示。具体的数据集特征和划分如下：

	字数	词数	句数	需要添加粗粒度表示的句数
训练集	1426813	704525	90000	14688
验证集	160161	86035	10000	1462
测试集	159862	84030	10000	1215

Table 2: 数据集划分

5 实现细节和参数设置

模型使用PCL-MedBERT作为基础编码器进行训练，同BERT (2018)相比该编码器使用医学文本语料预训练，可以提供更多医学文本特征。使用Adam (2014)优化器在训练中优化参数。为了防止模型的过拟合，我们在序列标注模块中的双向LSTM网络中添加了概率为0.4的Dropout，具体的超参数设置如表3。

词向量维度	768
输入最大长度	120
LSTM层数	2
LSTM表示的维度	200
表示组合器中的λ	1
词向量输入Dropout	0.5
LSTM层Dropout	0.4
判别器隐层维度	32
学习率	10-5
Batch_size	12

Table 3: 超参数设置

我们选择的基线方法包括pkuseg分词工具 (2019)、双向LSTM-CRF、BERT-LSTM-CRF和医学BERT-LSTM-CRF。

- pkuseg分词工具：pkuseg是一种支持多细分领域分词的分词工具。我们选用更适合于本文医学文本数据集的medicine细分领域模型进行实验。
- 双向LSTM-CRF：双向LSTM-CRF逐步读取输入中的字符，使用两层双向LSTM得到每个时间步所对应的隐藏状态表示，再使用CRF层进行全局归一化，捕捉字符标签之间的转移关系，对于每个输入字符进行分词标签的预测。
- BERT-LSTM-CRF：BERT-LSTM-CRF在谷歌发布的bert-base-chinese预训练模型基础上添加两层双向LSTM层和CRF层并进行finetune，给模型带来了更丰富的表示信息。
- 医学BERT-LSTM-CRF：医学BERT-LSTM-CRF相比BERT-LSTM-CRF，不使用在通用文本语料上预训练的bert-base-chinese而是选择使用在医学文本和医学问答数据上预训练得到的PCL-MedBERT，该编码器也与双编码器模型在实验中使用的编码器一致，可以提供更多来自医学领域的文本表示信息，而后续的序列标注方式同BERT-LSTM-CRF一致。

6 实验结果和分析

我们测试了各种基线方法以及本文提出的双编码器模型在测试集上的效果。如表4所示，我们提出的双编码器模型在各项指标上都超越了现有的通用中文分词工具以及现有的深度学习方法。本文提出的双编码器模型从两方面提升了医学文本分词的性能。一方面，通过判别器对于句子的判定给包含医学专用术语的文本提供了粗粒度的分词表示，提升了分词系统对于医学专用术语的分词能力。另一方面，由于模型给包含医学专用术语的文本单独提供了辅助编码器以粗粒度的分词，模型由于不完全负责医学专用术语的分词，也避免了这部分语料给模型主编码器的编码带来的干扰。

模型	P	R	F1
pkuseg	88.02	89.11	88.56
双向LSTM-CRF	89.25	90.97	90.10
BERT-LSTM-CRF	92.02	92.89	92.45
医学BERT-LSTM-CRF	92.56	93.12	92.84
双编码器	**92.96**	**93.19**	**93.08**

Table 4: 模型效果比较

然后，为了验证双编码器模型在需要粗粒度分词数据上的表现，我们将测试集中标记含有医学专用术语的数据提取了出来，形成了需要粗粒度分词的子测试集。我们使用之前实验中得到的基线方法及双编码器模型在总测试集上进行了多次测试，对每种方法的最优模型在子测试集上的性能取平均值，如表5所示。测试结果发现模型对于包含医学术语的中文分词能力上，相比不对医学术语文本提供粗粒度表示的医学BERT-LSTM-CRF模型有明显的提高。

另外，我们发现医学术语判别器在判断中无法达到完全正确是我们模型提升的瓶颈。由于模型在测试阶段需要对输入文本判断是否包含医学术语，模型中医学术语判别器的判别能力在F1指标上只能达到81.55%。如果直接将输入文本的标签输入模型而不使用医学术语判别器判别（即认为医学术语判别器的F1指标为100%），则在针对包含医学术语的子数据集上可以达到82.23，这也是当前结构模型在子数据集上的性能上限。虽然我们的判别器无法达到完全正确的效果，但是也证明了使用医学术语判别器判断是否包含医学专用术语并根据判别结果提供粗粒度表示对于提升医学分词系统能力上的作用。

为了探究双编码器模型中辅助编码器在分词粒度的区分上的帮助，我们将测试集划分为总测试集、测试集中包含医学专用术语的数据和测试集中不包含医学专用术语的数据三个部分。然后，在双编码器模型和最好的基线方法（医学BERT-LSTM-CRF）上选取最优模型分别针对这三部分数据统计了分词结果的平均词长，并将它们同测试集中的正确分词结果进行了对比，实验结果如表6。

通过实验发现，无论是基线模型还是双编码器模型都在整个测试集上的分词平均词长都低于正确分词结果，这主要因为在对于医学专用术语的切分上模型均出现一定数量的不必要切分，在包含医学专用术语的子测试集中两种模型分词结果的平均词长分别低于正确分词结

模型	P	R	F1	判别器F1
Jieba	58.95	78.29	67.26	-
双向LSTM-CRF	70.89	73.39	72.11	-
BERT-LSTM-CRF	75.46	79.48	77.42	-
医学BERT-LSTM-CRF	78.20	81.01	79.58	-
双编码器	**78.75**	**81.66**	**80.18**	**81.55**
双编码器（给定）	81.13	83.35	82.23	100

Table 5: 最优模型在包含医学术语的输入文本中的比较

模型	测试集	包含医学专用术语	不包含医学专用术语
医学BERT-LSTM-CRF	1.761	2.701	1.662
双编码器	1.767	2.827	1.661
正确分词结果	1.769	2.917	1.658

Table 6: 分词结果的平均词长

果0.09和0.216。但是，可以看出双编码器模型由于添加了辅助编码器，并使用这个辅助编码器产生医学专用术语需要的粗粒度表示，在整个测试集和包含医学专用术语的子测试集中得到的平均词长都要高于基线模型，并让平均词长更接近正确分词结果。在不包含医学专用术语的子测试集中，由于辅助编码器负责医学专用术语的粗粒度部分表示，这部分数据受到医学专用术语的粗粒度分词结果的干扰更小，让分词粒度有所下降，也更接近正确分词结果。

为了进一步探究判别器和辅助编码器在医学文本分词中对于模型的作用，我们进行了消融实验。实验结果如表7所示，我们可以看到：

1）在双编码器的基础上去除辅助编码器相当于在医学BERT-LSTM-CRF上添加了判断是否包含医学专用术语的多任务判断，实验可以发现，在任务中仅仅添加多任务监督对于整体分词效果的提升不明显，多任务训练提供的监督信息无法直接影响到分词表示上。

2）添加粗粒度分词表示则对分词效果提升更大，模型一方面通过辅助编码器提供的粗粒度分词表示丰富了输入文本表示，另一方面无需粗粒度分词的文本在训练中不会接触到医学专用术语，减少了相互之间的干扰，从而提升模型效果。

模型	F1
双编码器	93.08
-辅助编码器	92.86（-0.22）
-辅助编码器-判别器	92.84（-0.24）

Table 7: 消融实验

7 案例分析

为了展示双编码器模型在包含医学专用术语和不包含医学专用术语的两类文本中的作用，我们比较了部分案例在基线方法（医学BERT-LSTM-CRF）和双编码器模型运行的效果，如表8所示，“/”为分词标记。由表中案例可以看出：

1）针对包含医学专用术语的案例，双编码器模型将“中枢抑制性药物”看作了一个整体，而基线方法则将其分为了“中枢抑制性”和“药物”两个部分。由于双编码器模型中医学术语判别器对于输入文段中是否含有医学专用术语进行了判断，并根据结果通过辅助编码器给包含医学专用术语的输入提供了粗粒度表示。让模型对于医学专用术语的分词粒度与通用文本的分词粒度进行了区分。

2）由于双编码器模型中将不包含医学专用术语的医学通用文本同包含术语的文本分割开来，导致医学专用术语的粗粒度在训练中不会影响到数据集中的通用文本，排除了医学专用术语粗粒度切分的干扰后，模型对于不包含医学专用术语的医学通用文本的分词性能也有了明显

包含医学专用术语	案例	使用/中枢抑制性药物/对/智力/有/损害/等/。
	基线	使用/中枢抑制性/药物/对/智力/有/损害/等/。
	双编码器	使用/中枢抑制性药物/对/智力/有/损害/等/。
不包含医学专用术语	案例	束带/松解/束带/可/压迫/十二指肠/造成/反复/梗阻/,
	基线	束带/松解/束带/可/压迫/十二指肠/造成/反复梗阻/,
	双编码器	束带/松解/束带/可/压迫/十二指肠/造成/反复/梗阻/,

Table 8: 案例分析

提高。例如案例中，“反复梗阻”在基线方法中均被认为是一个整体，这显然是训练集中有大量包含粗粒度分词的医学专用术语在训练时给模型产生的影响。而当双编码器模型判断输入文本不包含医学专用术语时，模型只会使用产生通用分词粒度表示的主编码器提供的表示进行分词，如案例所示，“反复梗阻”在双编码器模型中被正确切分。

8 总结

我们在医学文本的中文分词任务中，针对医学文本中包含医学专用术语的问题，提出了双编码器模型。

由于在医学文本分词任务中，我们需要将文本中的医学专用术语看为一个分词整体，而医学专用术语往往包含各种前后缀，将其看作整体所需的分词粒度同数据集中的通用文本切分粒度明显不同。这种与通用文本分词所需粒度上的差异导致直接使用通用分词系统会带来两种分词粒度的相互干扰。在双编码器模型中，模型使用判别器对文本中是否包含医学专用术语进行判断，并对于包含医学专用术语的文本添加辅助编码器提供的粗粒度分词表示，这样不仅能为医学专用术语提供粗粒度分词表示，也不会因为医学专用术语的粗粒度切分影响到医学文本语料中通用文本表述的分词粒度，由此提升了整个医学文本分词系统的性能。

致谢

本文工作得到国家重点研发项目(2018AAA0102003)、国家自然科学基金(61876004)支持，特此致谢。

参考文献

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuan-Jing Huang. Gated recursive neural network for chinese word segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1744–1753, 2015.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuan-Jing Huang. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

R Luo, J Xu, Y Zhang, X Ren, and X Sun PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation 2019.

Ji Ma, Kuzman Ganchev, and David Weiss. State-of-the-art chinese word segmentation with bi-lstms. *arXiv preprint arXiv:1808.06511*, 2018.

Mairgup Mansur, Wenzhe Pei, and Baobao Chang. Feature-based neural language model and chinese word segmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1271–1277, 2013.

Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, 2004.

Wenzhe Pei, Tao Ge, and Baobao Chang. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303, 2014.

Fuchun Peng. Chinese segmentation and new word detection using conditional random fields. 2004.

Huihsin Tseng, Pi-Chuan Chang, Galen Andrew, Dan Jurafsky, and Christopher D Manning. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Jingkang Wang, Jianing Zhou, Jie Zhou, and Gongshen Liu. Multiple character embeddings for chinese word segmentation. *arXiv preprint arXiv:1808.04963*, 2018.

Jingjing Xu and Xu Sun. Dependency-based gated recursive neural network for chinese word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–572, 2016.

Nianwen Xue. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48, 2003.

Huan Zhang, Yuan Zong, Baobao Chang, Zhifang Sui, Hongying Zan, and Kunli Zhang. 面向医学文本处理的医学实体标注规范(medical entity annotation standard for medical text processing). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 561–571, 2020.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. Neural networks incorporating dictionaries for chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. Deep learning for chinese word segmentation and pos tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 647–657, 2013.

张立邦, 关毅, and 杨锦峰. 基于无监督学习的中文电子病历分词. 智能计算机与应用, 2014.

王莉军, 周越, 桂婕, and 翟云 基于BiLSTM-CRF的中医文言文文献分词模型研究. 计算机应用研究, v.37;No.349(11):165-168+173, 2020.

王若佳, 赵常煜, and 王继民. 中文电子病历的分词及实体识别研究. 图书情报工作, 63(2):34–42, 2019.

黄丹丹and 郭玉翠. 融合attention机制的bi-lstm-crf中文分词模型. 软件, 39(10):268–274, 2018.