计算机科学
COMPUTER SCIENCE

中文工艺规范文本分词语料的构建与研究

王裴岩, 张莹欣, 付小强, 陈佳欣, 徐楠, 蔡东风

引用本文

王裴岩, 张莹欣, 付小强, 陈佳欣, 徐楠, 蔡东风. 中文工艺规范文本分词语料的构建与研究[J]. 计算机科学, 2023, 50(11A): 221200070-6.

WANG Peiyan, ZHANG Yingxin, FU Xiaoqiang, CHEN Jiaxin, XU Nan, CAI Dongfeng. Construction and Research of Chinese Word Segmentation Corpus of Process Specification Text [J]. Computer Science, 2023, 50(11A): 221200070-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200070

中文工艺规范文本分词语料的构建与研究

王裴岩[1] 张莹欣[1] 付小强[2] 陈佳欣[1] 徐 楠[1] 蔡东风[1]
1 沈阳航空航天大学人机智能研究中心 沈阳 110136
2 中国商飞上海飞机制造有限公司航空制造技术研究所 上海 201324

摘 要 中文分词是处理工艺规范文本的一项基本任务,并且在工艺知识图谱与智能问答等下游任务中发挥着重要作用。工艺规范文本分词面临的一个挑战是缺乏高质量标注的语料,特别是面向术语、名词短语、工艺参数、数量词等特殊语言现象的分词规范。文中面向工艺规范文本制定了专用分词规范,收集并标注了一个中文工艺规范文本分词语料(WS-MPST),含11900个句子与255160个词,4名标注者分词标注一致性达95.25%。在WS-MPST语料上对著名的BiLSTM-CRF与BERT-CRF模型进行了对比实验,F1值分别达到92.61%与93.69%。实验结果表明,构建专用的工艺规范分词语料是必要的。对实验结果的深入分析揭示了未登录词与中文非中文字符混合构成的词是工艺规范文本分词的难点,也为今后工艺规范文本及相关领域的分词研究提供了一定的指导。

关键词:中文分词;工艺规范文本;分词规范;分词语料;分词模型
中图法分类号 TP391

Construction and Research of Chinese Word Segmentation Corpus of Process Specification Text

WANG Peiyan[1],ZHANG Yingxin[1],FU Xiaoqiang[2],CHEN Jiaxin[1],XU Nan[1] and CAI Dongfeng[1]
1 Human-Computer Intelligence Research Center,Shenyang Aerospace University,Shenyang 110136,China
2 Aviation Manufacturing Technology Research Institute,COMAC Shanghai Aircraft Manufacturing,Shanghai 201324,China

Abstract Chinese word segmentation is a basic task for process specification text processing,which has a critical impact on downstream tasks such as process knowledge graphs and intelligent Q&A systems. One of the challenges faced by word segmentation of process specification texts is the lack of high-quality annotated corpus,especially word segmentation specifications for special language phenomena such as terms,noun phrases,process parameters,and quantifiers. This paper formulates a special word segmentation specification for the process specification text,collects and annotates a word segmentation corpus for Chinese process specification text(WS-MPST),including 11900 sentences and 255160 words,and the consistency of word segmentation by 4 annotators achieves 95.25%. On the WS-MPST corpus,the famous BiLSTM-CRF and BERT-CRF models are tested,and the F1 values achieves 92.61% and 93.69% respectively. Experimental results show that it is necessary to construct a special word segmentation corpus for process specification test. The in-depth analysis of experimental results reveals that the out-of-vocabularywords and the words which contain Chinese and non-Chinese characters are difficultto segment in process specification texts,which provides some guidance for future word segmentation research in process specification texts and related fields.

Keywords Chinese word segmentation,Process specification text,Word segmentation specification,Word segmentation corpus,Word segmentation model

1 引言

工艺规范是对工艺过程中有关技术要求所做的一系列规定,主要包括工艺参数和工艺条件[1]。工艺规范是工艺技术工作行为的规定,也是工艺审查与制造质量控制的依据,应用于设计、制造与质量管理的全流程中。在智能制造与数据经济的时代背景下,制造业企业越来越重视对非结构化文本数据的开发与利用。中文分词是将文本(即字符序列)分割成单词,是中文文本处理的关键组件之一。工艺规范文本的分词对工艺知识图谱[2]与问答系统[3]等智能应用具有重要的作用。

目前,面向制造领域,特别是面向工艺规范文本的分词语料、方法与工具非常匮乏,研究非常不足。按照制造领域语言特点及工艺文本处理需求所研究制定的分词规范与标注语料未见公开报道,严重制约了该领域分词算法的研究与应用。现有制造领域分词应用,通过加入领域词典或者构建小规模语料的方式,解决通用领域分词工具在制造领域面临的未登录词与切分歧义等问题。例如,Zhu等[3]在机械智能制造知识问答系统的数据预处理阶段使用了NLPIR分词系统[1)],通过加入领域词典的方法补充领域词,解决NLPIR未登录词问题。Gu等[2]在构建工艺知识图谱中应用了分词与词性标注,

1) http://ictclas.nlpir.org/

基金项目:辽宁省应用基础研究计划(2022JH2/101300248)
This work was supported by the Applied Basic Research Program of Liaoning Province (2022JH2/101300248).
通信作者:王裴岩(wangpy@sau.edu.cn)

针对专有名词或术语在通用中文语料数据集的训练下通常会被统一认为是普通名词的问题，构建了干预语料，面向工艺知识图谱构建需求细分了词类，从而更好地定义实体类语义和关系类语义。Chen 等[4]指出工艺语义信息蕴含在非结构化的装配工艺文档中，提出了长短时间记忆网络的提取方法。为训练长短时间记忆网络，他们构建了 1139 句 12256 词的分词及词类标注语料用于构建装配工艺知识图谱，并指出扩大数据集的语料量可以提高装配实体和关系的语义识别率，保证装配知识图谱的表达正确率。这也反映出扩大语料规模有助于获得更为广泛的工艺规范文本分词语义信息，提高分词的准确性。

为了对工艺规范文本进行分词，研究员面临着一个挑战，即缺乏专门的分词标注语料用于模型的训练，或是标注语料数据量太小。虽然有一种可能的解决方案是将在通用领域训练的模型应用到工艺规范文本中，但这些模型没有良好的性能。工艺规范中许多领域术语很少在通用领域出现。大量研究[5-6]表明采用词表的方法无法处理未登录词问题。另外，制造领域也有区别于通用领域的语言现象。例如，文本中包含了大量的文件编号、零件编号、材料编号、技术参数、简称与缩略语等非中文字符。这类语言现象在中文分词评测 SIGHAN Bakeoff 2005 语料[7]中极少出现，北大分词语料(PKU)为 1.9‰，微软亚洲研究院分词语料(MSR)仅为 0.076‰。这种现象一方面可作为切词的自然标记，另一方面，非中文字串的多样性也造成了分词语境的复杂性。通用领域语料所训练的分词模型与分词工具较难处理此类问题。这表明了面向制造领域制定专门的分词标注构建分词语料的必要性。

从语料及分词方法研究方面来看，中文分词评测 SIGHAN Bakeoff 2005 语料与 CTB(Penn Chinese Treebank)[8]语料是最常用的语料，已经成为了分词算法研究的基准数据。基于上述语料的分词算法的效果不断提升，PKU 语料上 F1 值已经能够达到 96%以上[9-10]，MSR 语料上 F1 达到了 98%以上[9-10]。上述语料的数据来源主要是新闻及社交媒体文本，在上述语料上开发的分词算法不能适应专门领域的分词需要，效果表现较差。Liu 等[11]收集并标注了中文电子病历分词及实体识别语料，其实验表明在 CTB 上训练的模型在中文电子病例上 F1 值仅达到 77.22%，而在中文电子病例语料上训练的模型达到 99.03%的 F1 值，表明了专门构建语料的必要性。Liu 等[12]则研究了中文分词领域适应问题，构建了中文小说 Zhuxian 语料。Qiu 等[13]专门研究了中文小说文本分词问题，指出中文分词在新闻数据上非常准确，但在其他领域，如科学和文学领域，准确性则大幅下降。有研究表明，科技领域分词的困难主要是领域术语问题，如专利领域分词针对术语进行前处理或后处理等[14-16]。同上述领域及文本相比，制造领域中文分词的研究十分不足，未见相关研究与文献发表。与操作工艺文档、医疗和专利等专业领域相比，除领域术语外，名词短语、工艺参数、数量词等表现出了特有的构词现象。

面对上述问题，本文制定了面向中文工艺规范文本的分词规范，收集并标注了一个新的中文工艺规范文本分词语料，主要贡献包括：

1)制定了中文工艺规范文本分词标注规范：面向制造领域专业名称与术语、名词性短语、工艺参数、数量词、编号、章节号、图注与表注这 7 类特有语言现象，制定了专门的分词规范，为制造领域和相关领域文本分词规范的制定提供了依据与参考。

2)人工标注了分词语料：该语料含 11900 句，485015 字符，255160 词。通过实验，验证了所构建语料能够支撑基于双向长短时记忆网络与条件随机场(BiLSTM-CRF)的神经网络分词模型以及预训练语言模型的微调模型(BERT-CRF)的训练，分词 F1 值分别达到 92.61%与 93.69%。并且，在通用领域数据集 Sighan 2005 backoff 分词评测语料 PKU 与 MSR 上训练的模型在制造领域上表现较差，这证实了工艺规范文本分词构建语料及训练模型的必要性。

3)分析了未登录词对工艺规范文本分词的影响：通过对语料及模型分词错误的分析，发现 4 字词切分精度较低。相比于通用领域，工艺规范文本的未登录词更不容易被正确切分。BERT-CRF 模型未登录词的召回率为 66.18%，显著低于其在 PKU 与 MSR 上的86.76%与 86.67%，未登录词错误是词典词错误的 6～7 倍。

4)分析了工艺规范特有的工艺参数、编号、章节号等非中文字符词分词效果及其对句内其他词的影响：发现由中文及非中文字符混合构成的词具有相对较高的错误率(16.67%)，是工艺规范文本分词的难点。

本文第 1 章为引言；第 2 章介绍分词标准与语料构建过程，并对语料做了统计；第 3 章介绍了实验的分词模型 BiLSTM-CRF 与 BERT-CRF；第 4 章报告了实验结果，并详细分析了未登录词与不同构词的分词效果；最后总结全文并展望未来。

2 分词语料库的构建

2.1 数据收集

本文收集了互联网公开制造领域国家标准工艺规范文档 400 份，涉及装配、复材加工、普通机械加工与数控加工工艺领域。每份文档包括：适用范围、引用文件、材料控制、设备控制、技术控制、程序控制、维护控制与质量要求 8 个章节。从材料控制、设备控制、技术控制、程序控制、维护控制与质量要求 6 个涉及工艺参数和工艺条件要求的章节中抽取语句，过滤掉涉及具体型号与企业的敏感信息。被抽取的语句至少包含 10 个中文字符，从而保证能够有足够的中文词信息，获得 11900 句语料标注分词。

2.2 工艺规范文本分词规范

为确保标注语料的标注质量，本文邀请 4 名制造领域专业人员制定分词标准，并标注语料。该 4 名标注人员包括编制和应用工艺规范的两类人员，能够同时兼顾编制与应用两方面制定标注规范，具体为 2 名工艺规范编制人员和 2 名工艺员。分词标准基于《信息处理用现代汉语分词规范》[17](简称《规范》)，并面向工艺规范语言特点予以修改和扩展。领域术语参考全国科学技术名词审定委员会的“术语在线(termonline)”网站[1)]。

针对专业名词与术语、名词性短语、工艺参数、编号、章节号、图注与表注等特殊语言现象，做出了区别于《规范》的专门

1) https://www.termonline.cn/

规定。具体包括以下 7 类：

1）专有名词与术语：术语在线包含的词或是表示一个制造领域技术、产品、工具、材料等概念的词。例如“铆钉”“密封剂”“铣削”“树脂”“脱模剂”。需要特别指出的是，如果专有名称与术语由英文及中文字符组合构成，则作为一个词不切分，如“Tedlar 膜”“MC 涂层”“AWG10 导线”。

2）专有名词与术语构成的名词短语：表示一个制造领域技术产品工具材料等概念的短语，切分后会改变语义连贯性或改变语义内涵，故不予切分，便于后续实体识别与翻译等应用作为独立单元处理，如“抽芯铆钉”“盲铆钉”“聚硫醚密封剂”“电磁铆接设备”“压敏胶粘剂”。

3）工艺参数：工艺参数等数值与单位组合成词不切分，例如“310 kPa±34 kPa”“－0.06 MPa”“－0.75 mm”“1/4”等。中文单位（如“秒”与“英寸”等）也不切分，如“4 小时”“28 天”“10 秒钟”“0.5 英寸”。由“－”或“～”等符号连接代表区间的参数作为一个整体不切分，如“≤0.7 mm”“7.2-7.8”“13 毫米～25 毫米”等。此条规定与《规范》有所区别，其原因是：制造领域工艺参数是最为重要的信息之一，不切分能够保证工艺参数语义的完整性。

4）数量词：数词和量词组成的数量词不切分，如“一层”“一块”“3 支”。此条规定也与《规范》不同，在制造领域，数量也是一类重要技术要求，可视为一种指标或参数，因此采用了与工艺参数相同的规定。

5）编号：文件、材料、设备、工具等编号不切分，如“REM7511-61”与“TEE-WL-629”。

6）章节号：工艺规范内引用文件的章节号不切分，如“6.2.2.3 节”，“1.2 节”。

7）图注与表注：引用的图注与表注组合成词不切分，如“表 3-1”“图 2”“附图 2”。

除上述 7 条规定外，其他类词延用《规范》内的要求，这样能够使得语料中的通用词与通用领域语料一致，从而便于通用领域词向量与预训练语言模型等在制造领域中的迁移应用。

2.3 语料标注

工艺规范文本分词语料的标注过程分为 3 个阶段。第一阶段标注的目的是：标注者制定与理解分词规范，进行试标注，验证与迭代分词标准。第一阶段，均匀随机抽取 200 句作为标注语料，如果分词规范修改，再次抽取 200 句予以验证。当分词规范固定后，开展第二阶段标注。第二阶段标注的目的是验证标注者对于分词规范理解的一致性，抽取 200 句同时开展标注，之后计算一致性。标注者的分词标注一致性采用 F1 值评价[18]。具体的方法是将一位标注者（$A1$）的标记结果视为标准答案，并计算另一位标注者（$A2$）的标注结果的 $F1$ 值（计算方法如式 3）。当 4 位标注者的评价 F1 值达到 95%以上时，开展第三阶段标注。第三阶段，将剩余语料分为 2 份，每份语料由两位标注者同时开展标注工作，完成全部语料的标注。最终，4 位标注者的平均标注一致性为 95.25%。

$$P=\frac{A1\text{ 和 }A2\text{ 标注一致的词数}}{A2\text{ 标注语料的词数}}\times 100\% \tag{1}$$

$$R=\frac{A1\text{ 和 }A2\text{ 标注一致词数}}{A1\text{ 标注语料的词数}}\times 100\% \tag{2}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{3}$$

2.4 语料统计与分析

本文将标注的分词语料命名为工艺规范文本分词语料（Word Segmentation corpus of Manufacturing Process Specification Text，WS-MPST）。表 1 列出了 WS-MPST 语料的句子长度、词长、字符数、词数等统计量。

表 1 WS-MPST 语料统计

Table 1 Statistics of WS-MPST

	数量	长度			
		平均值	标准差	最大值	最小值
字符	485 015	—	—	—	—
词	255 160	1.90	1.28	20	1
词 C1	50 238	2.15	2.25	20	1
词 C2	1 427	3.62	1.76	13	2
词 C3	203 495	1.83	0.87	12	1
句子	11 900	40.76	30.09	394	3

注：词 C1 为由非中文字符构成的词；词 C2 为由中文及非中文字符构成的词；词 C3 为全中文字符构成的词。

表 1 还专门统计了非中文字符构成的词（词 C1）、由中文及非中文字符构成的词（词 C2）、全中文字符构成的词（词 C3）的情况。此三类词分别对应了第 2.2 分词规范的不同要求，特别是词 C1 与词 C2 针对“3）－7）”工艺规范特有的构词现象。词 C1 与词 C2 的数量分别为 50 238 与 1 427，占比分别为 19.69% 与 0.56%，显著高于 PKU（1.9‰）与 MSR（0.076‰）。表 2 列出了 WS-MPST 分词语句示例。

表 2 WS-MPST 分词示例

Table 2 Word segmentation examples from WS-MPST

句子示例	说明
用 ‖ 最大 ‖ **0.086 毫米** ‖（‖ **0.0034 inch** ‖）‖ 的 ‖ 钢丝 ‖ 制 ‖ 的 ‖ **不锈钢丝刷** ‖，‖ 刷 ‖ 整个 ‖ 接触 ‖ 表面	工艺参数“0.086 毫米”与“0.0034inch”； 名词性短语“不锈钢丝刷”。
在 ‖ 每个 ‖ 接触 ‖ 的 ‖ 表面 ‖ 上 ‖ 均匀 ‖ 地 ‖ 涂 ‖ 一层 ‖ **PMT-AL-314** ‖ **防粘结石油基润滑剂**	工艺材料编号“PMT-AL-314”； 名词性短语“防粘结石油基润滑剂”。
7.3.7 ‖ 燃油 ‖ 传输 ‖ 槽口 ‖ 平底 ‖ 压印 ‖ 的 ‖ 典型 ‖ 方式 ‖ 参见 ‖ **图 7-5** ‖（‖ **机翼长桁** ‖）‖。	章节号“7.3.7”； 图注“图 7-5”； 名词性短语“机翼长桁”。

本文还统计了词长的分布，并且与 MSR 和 PKU 语料进行了比较，如图 1 所示，WS-MPST 与 MSR 及 PKU 词长分布相近，大部分词长度为 1 和 2。其原因是通用领域词按照《规范》要求切分，因此 WS-MPST 中通用领域词的长度与 MSR 及 PKU 通用领域语料一致。WS-MPST 中长度大于 3 的词的比例相比其他语料较高，达到 17.8%，PKU 为 9.1%，MSR 为 10.8%。这些较长词大部分为专有名词组成的名词性短语、工艺参数与编号等，长词势必会造成分词的困难。

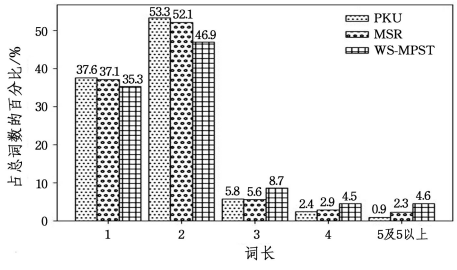

图 1 WS-MPST，MSR 与 PKU 语料词长分布图

Fig. 1 Diagram of word length distributions of WS-MPST, MSR and PKU

3 分词模型简介

本文遵循中文分词研究的字位序列标注方法[5-6]，字位标记采用“BIES”方案。“B”表示词首字；“I”表示词中字；“E”表示词尾字；“S”表示单字词。本文选择了神经网络模型与预训练语言模型微调2种不同方式的2个众所周知的模型：BiLSTM-CRF与BERT-CRF。选取这两个模型的原因是想验证WS-MPST是否能够支撑不同方式的分词模型训练，而且这两个模型也能够验证基于通用领域预训练语言模型微调是否能够适用于制造领域分词。

BiLSTM-CRF[19]是神经网络编码器-解码器结构的模型，其中BiLSTM作为编码器，CRF[20]作为解码器，是分词[21-22]等序列标注任务最常用的神经网络模型。BERT-CRF是基于预训练语言模型BERT微调的分词模型，BERT[23]是应用最为广泛的预训练语言模型，解码器也使用了CRF。

4 实验与分析

4.1 实验设置

将WS-MPST语料分为训练集、验证集与测试集，分别用于模型训练、模型选择与模型测试。训练集、验证集与测试集比例为8∶1∶1。表3列出了训练集、验证集与测试集的统计信息。另外，本文还使用了通用领域文本的中文分词基准数据集Sighan 2005 backoff分词评测语料PKU与MSR。

表3 数据集统计

Table 3 Dataset statistics

	训练集	验证集	测试集
字符数	385066	50035	49914
词数	202712	26295	26153
词(C1)	39883	5224	5131
词(C2)	1120	151	156
词(C3)	161709	20920	20866
句子数	9520	1190	1190

使用中文分词任务中常用的评价指标精准率P(Precision)、召回率R(Recall)、F1值与未登录词的召回率(Roov)作为模型的评价指标。模型超参数设置如表4所列。BiLSTM-CRF的嵌入层维度(Embedding Size)是300维。BiLSTM隐藏层维度(Hidden Size)是256维度。训练批量大小(Batch Size)为32，训练周期(Epoch)设为30，学习率(Learning Rate)设为1×10^{-3}，暂退率(Dropout Rate)设置为0.2。预训练模型采用bert-base-chinese[1)]，在其基础上作模型微调，训练批量大小为4；训练周期为20，学习率设为1×10^{-5}，暂退率为0.1。

表4 模型超参数设置

Table 4 Model hyper-parameters settings

参数	BiLSTM-CRF	BERT-CRF
Embedding Size	300	768
Batch Size	32	4
Epoch	30	20
Learning Rate	1×10^{-3}	1×10^{-5}
Hidden Size	256	768
Dropout Rate	0.2	0.1

4.2 WS-MPST分词实验

表5列出了WS-MPST分词实验结果，实验分为3种不同设置，即WS-MPST，PKU，MSR。其中，PKU与MSR表示分别在PKU及MSR训练集上进行训练模型，在WS-MPST测试集测试模型。WS-MPST表示在WS-MPST训练集上训练模型，在测试集上测试模型。由表5可见，在WS-MPST上训练的BiLSTM-CRF与BERT-CRF的F1值分别达到了92.61%与93.69%。这表明所构建的WS-MPST语料能够支撑神经网络模型与预训练语言模型微调2种不同方式的分词模型训练，也验证了基于通用领域预训练语言模型微调方式能够适用于工艺规范文本分词。与PKU及MSR上训练的模型相比，WS-MPST上训练的所有模型各项指标都有明显的提升，特别是PKU与MSR训练的BiLSTM-CRF模型仅有38.19%及35.55%的F1值。较差的结果证实了分词规范及语料在通用领域和工艺规范文本之间的巨大差距，表明了从通用领域到工艺规范文本的转移应用的挑战，也进一步强调了在工艺规范文本提出分词规范构建分词语料的必要性。

表5 WS-MPST分词实验结果

Table 5 Word segmentation experiment results on WS-MPST

(单位：%)

方法	WS-MPST					PKU					MSR				
	P	*R*	*F1*	*OOV*	R_{OOV}	*P*	*R*	*F1*	*OOV*	R_{OOV}	*P*	*R*	*F1*	*OOV*	R_{OOV}
BiLSTM-CRF	92.43	92.80	92.61	4.16	58.55	32.54	46.21	38.19	—	—	30.05	43.51	35.55	—	—
BERT-CRF	93.79	93.59	93.69	4.16	66.18	78.59	86.74	82.46	—	—	71.98	82.51	76.89	—	—

表5同时还展示了WS-MPST语料的未登录词率(OOV)及未登录词召回率(R_{OOV})。WS-MPST语料测试集的未登录词率为4.16%，低于PKU的5.8%[5]，高于MSR的2.6%[5]。未登录词的比率并不是很高。然而，从未登录词召回率来看，WS-MPST的召回率并不理想，BiLSTM-CRF模型为58.55%；BERT-CRF模型为66.18%。在Tian等[10]的研究中，PKU与MSR语料上，BiLSTM-CRF模型未登录词召回率分别达到56.80%与68.75%；BERT-CRF模型达到86.76%与86.67%。BERT-CRF模型上，WS-MPST语料上的未登录词召回率与PKU、MSR相差20%。这结果表明WS-MPST的未登录词更难识别。从分词错误角度来分析，BiLSTM-CRF词典词的错误率(错误词数/总词数)为5.67%；未登录词的错误率为42.46%。BERT-CRF词典词的错误率为5.18%；未登录词的错误率为34.65%。WS-MPST未登录词的错误率为词典词错误率的6～7倍。

表6列出了BiLSTM-CRF与BERT-CRF在不同构词上(词C1、词C2与词C3)的分词错误率与错词比例。分词错误率是指某类错词占该类词总量的百分比。错词比例是指某类错词占全部错词的百分比。为了便于分析，表6列出了测试集中各类词的占比。可以发现：词C2与词C3的错词比例比

1) https://huggingface.co/bert-base-chinese

各自的占比要高，而词 C1 相对要低；词 C1 的错误率最低，而词 C2 最高。词 C1 是全部由非中文字符构词的词，包括零件、材料与文件等的编号，构词规律简单，并且有自然的切分边界（中文与非中文字符边界）。这部分词在测试集中的占比达到了 19.62%，从前述的错误率及错词比例来看，分词难度不高。词 C2 是由中文及非中文字符混合构成，虽然占比不高仅为 0.60%，但从错误率和错词比例来看，此类词较难切分。并且词 C1 与词 C3 占比相对较高，中文与非中文字符间自然切分规律较强，对词 C2 也造成了影响。表 7 从词表词与未登录词两方面进一步展示了各类词的分词错误率，强调了词 C2 的分词难点。此类词多为数值与单位组合，代表了技术参数，属于工艺参数要求的重要部分。本文认为此类词是工艺规范文本分词研究的重点之一。

表 6　BiLSTM-CRF 与 BERT-CRF 在不同构词上的分词错误率与错词比例

Table 6　Word segmentation error rates and wrong word ratios of BiLSTM-CRF and BERT-CRF on different word formations

（单位：%）

词类型	比例	BiLSTM-CRF		BERT-CRF	
		错词比例	错误率	错词比例	错误率
词(C1)	19.62	11.15	4.09	11.76	3.84
词(C2)	0.60	1.75	21.15	1.55	16.67
词(C3)	79.78	87.10	7.86	86.69	6.96

表 7　BiLSTM-CRF 与 BERT-CRF 在不同构词上的词典词与未登录词分词错误率

Table 7　In-vocabulary and out-of-vocabulary word error rates of BiLSTM-CRF and BERT-CRF on different word formation word

（单位：%）

词类型	BiLSTM-CRF		BERT-CRF	
	词典词	未登录词	词典词	未登录词
词(C1)	3.08	25.76	2.88	24.45
词(C2)	10.26	53.85	6.84	46.15
词(C3)	6.28	46.59	5.73	36.95

图 2 给出了 BiLSTM-CRF 与 BERT-CRF 模型词长错误率分布，可见词长为 4 及词长大于或等于 5 的的词错误率较高。对长词的识别一直是分词的难点[24]。首先，词长大于 4 的词多为未登录词，占未登录词的 47.89%。其次，结合表 2 词长度分布，词 C2 的平均值为 3.6，标准差为 1.76。因此，词长大于 4 的词多为由中文及非中文字符混合构成的词，也表明了此类词是制造领域中文分词的难点。

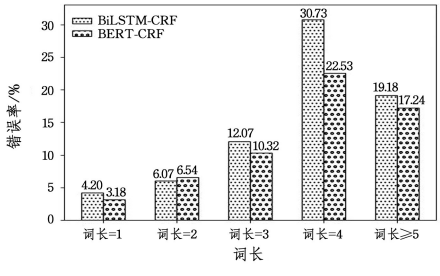

图 2　词长错误率分布图

Fig. 2　Diagram of error rate distribution on word length

结束语　本文构建了中文工艺规范文本分词语料 WS-MPST，含 11 900 句，255 160 词。在 WS-MPST 上，测试了 BiLSTM-CRF 与 BERT-CRF 分词模型。在 WS-MPST 上的实验结果表明，在通用领域数据集（PKU 与 MSR）上训练的模型在工艺规范文本中的效果较差，这证实了面向工艺规范文本分词问题构建专门分词规范及构建语料的必要性。此外，通过对分词结果的详细分析发现，由中文及非中文字符混合构成的词具有较高的错误率，是工艺规范分词的难点。由于此类词多为数值与单位组合，代表了技术参数，属于工艺参数要求的重要部分，因此，本文认为中文及非中文字符混合构成的词的处理是工艺规范文本分词研究的难点和重点。

在未来工作中，可以针对词 C1 与词 C2 类开展研究。词 C1 多为零件、材料与文件等编号，可以尝试利用企业文件编号命名规则构建识别规则[25]。词 C2 多为数值与单位组合，可利用常用计量单位词表，构建含触发词的识别规则。通过上述规则，预识别词 C1 与词 C2，并设计与研究能够利用预识别信息的分词模型。同时，对 BERT-CRF 的测试，表明预训练语言模型的微调方法能够用于工艺规范分词任务。但由于没有工艺规范文本的预训练模型，本文采用的是通用领域的模型。因此，收集大量的领域文本构建领域内模型，再应用于工艺规范文本分词，也是后续可开展的工作。此外，本文未来将公开软件工具和工艺规范文本分词语料库。

参 考 文 献

[1] China National Committee for Terminology in Science and Technology. Mechanical Engineering Terms (Second Edition) [M]. Beijing: Science Press, 2021.

[2] GU X H, BAO J S, LV C F. Assembly semantic information modeling based on knowledge graph[J]. Aeronautical Manufacturing' Technology, 2021, 64(4): 74-81.

[3] ZHU J N, LIANG Y Q, GU F, et al. Design of knowledge question-answering system for mechanical intelligentmanufacturing based on deep learning[J]. Computer Integrated Making System, 2019, 25(5): 1161-1168.

[4] CHEN Z Y, BAO J S, ZHENG X H, et al. Semantic recognition method of assembly process based on LSTM[J]. Computer Integrated Making System, 2021, 27(6): 1583-1593.

[5] HUANG C L, ZHAO H. Chinese Word Segmentation: A Decade Review[J]. Journal of Chinese Information Processing, 2007 (3): 8-19.

[6] ZHAO H, CAI D, HUANG C L, et al. Chinese Word Segmentation: Another Decade Review (2007-2017) [J]. arXiv: 1901.06079, 2019.

[7] EMERSON T. The Second International Chinese Word Segmentation Bakeoff[C]// Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Jeju Island, Korea, 2005: 123-133.

[8] XUE N W, XIA F, CHIOU F D, et al. The Penn Chinese Treebank: PhraseStructure Annotation of a Large Corpus[J]. Natural Language Engineering, 2005, 11(2): 207.

[9] HUANG K Y, HUANG D G, LIU Z, et al. A Joint Multiple Criteria Model in Transfer Learning for Cross-domain Chinese Word Segmentation[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 3873-3882.

[10] TIAN Y, SONG Y, XIA F, et al. Improving Chinese Word Segmentation with Wordhood Memory Networks[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[11] LIU Y, TIAN Y, CHANG T H, et al. Exploring Word Segmentation and Medical Concept Recognition for Chinese Medical Texts[C]// Proceedings of the 20th Workshop on Biomedical Language Processing. 2021: 213-220.

[12] LIU Y, ZHANG Y. Unsupervised Domain Adaptation for Joint Segmentation and POS-Tagging[C]// Proceedings of CoLING 2012. 2012: 745-754.

[13] QIU L K, ZHANG Y. Word Segmentation for Chinese Novels [C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015: 2440-2446.

[14] ZHANG G P, LIU D S, YIN B S, et al. Research on Chinese Word Segmentation for Patent Documents[J]. Journal of Chinese Information Processing, 2010, 24(3): 112-116.

[15] ZHANG J, ZHANG H C, ZHAI D S, et al. Research of the Word Segmentation for Chinese Patent Claims[J]. New Technology of Library and Information Service, 2014(9): 91-98.

[16] YUE J Y, XU J A, ZHANG Y J. Chinese Word Segmentation for Patent Documents[J]. Journal of Peking University, 2013, 49(1): 159-164.

[17] GB/T 13715-1992, Contemporary Chinese language word segmentation specification for information processing[S]. Beijing: China Standard Press, 1992.

[18] HRIPCSAK G, ROTHSCHILD A. Agreement, the F-measure, and Reliability in Information Retrieval[J]. Journal of the American medical informatics association, 2005, 12(3): 296-298.

[19] HUANG Z, WEI X, KAI Y. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv: 1508.01991, 2015.

[20] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of ICML'01. 2001: 282-289.

[21] MA J, GANCHEV K, WEISS D, et al. State-of-the-art Chinese Word Segmentation with Bi-LSTMs[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4902-4908.

[22] GONG J, CHEN X, GUI T, et al. Switch-LSTMs for Multi-Criteria Chinese Word Segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 6457-6464.

[23] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.

[24] SUN X, ZHANG Y Z, MATSUZKI T, et al. A discriminative latent variable Chinese segmenter with hybrid word/character information[C]// Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009: 56-64.

[25] GB/T 24735-2009. Numbering Method for Machine-Building Technological Documentation [S]. Beijing: China Standard Press, 2009.

WANG Peiyan, born in 1983, Ph.D, senior engineer, is a member of China Computer Federation. His main research interests include natural language processing, machine learning and knowledge engineering.