面向国防科技领域的技术和术语识别方法研究

冯鸾鸾 李军辉 李培峰 朱巧明
(苏州大学计算机科学与技术学院 江苏 苏州 215006)
(江苏省计算机信息技术处理重点实验室 江苏 苏州 215006)

摘 要 随着自然语言处理技术的发展,人们越来越重视构建面向国防科技领域的知识图谱。而面向国防科技领域的技术和术语识别是构建该领域技术知识图谱的基础。文中基于该领域的语料库,在技术和术语识别的任务上,探索了子词单元在传统序列标注 Bi-LSTM+CRF 模型上的应用。此外,针对任务的特点,提出了适用于技术和术语识别的语言学特征。基于该领域的语料库,实验结果表明技术和术语识别的 $F1$ 值达到了 71.80%,较基准系统提升了 3.04%,能够较好地识别出面向国防科技领域的技术和术语。同时,所提方法也优于基于 BERT 模型的技术术语识别方法。

关键词 面向国防科技领域,技术和术语,子词,Bi-LSTM+CRF 模型,语言学特征
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/jsjkx.190300069

Technology and Terminology Detection Oriented National Defense Science

FENG Luan-luan LI Jun-hui LI Pei-feng ZHU Qiao-ming
(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China)
(Provincial Key Laboratory for Computer Information Processing Technology, Suzhou, Jiangsu 215006, China)

Abstract With the rapid development of natural language processing, constructing oriented national defense science (ONDS) technology knowledge base has received more and more attention. The identification of technology and terminology is fundamental for constructing ONDS technology knowledge base. To recognize technology and terminology, this paper explored the application of subwords in the traditional Bi-LSTM+CRF sequence labeling model. In addition, this paper proposed linguistic features to boost the performance. Experimental results on the annotated dataset show that the proposed approach achieves 71.8% $F1$ scores, with improvement of 3.04% over the baseline system, indicating the effectiveness of the proposed approach in recognizing ONDS technology and terminology. Meanwhile, it also outperforms BERT-driven models in recognizing technology and terminology.

Keywords Oriented national defense science, Technology and terminology, Subwords, Bi-LSTM+CRF model, Linguistic features

1 引言

随着互联网海量信息的不断增长,从大数据中挖掘有价值的信息并将其应用于国防建设是必然的趋势。在互联网上,存在海量的文献和科技信息,可以从中得到高价值情报。研究如何从大数据中抽取世界各国的国防相关技术及其研发和应用信息,有助于把握国防技术发展的态势,为我国国防建设服务。

在国防科技领域,技术和术语不同于传统意义上的实体。例如,在例 1 中,Synthetic aperture radar 及其缩略语 SAR 都可以认为是特殊的实体,即本文的技术;airborne SAR system 是一个名词短语,同时也是一项技术。面向国防科技领域的技术和术语相对于传统实体而言通常较长,有的可能同时包含形容词和连词,并且一项技术或术语可能有多种不同的表现形式,比如简写或者首字母缩略词。

例 1 The Synthetic aperture radar (SAR) technology has been developed in China since 1970s, and the first airborne SAR system was established in 1979 and obtained multiple SAR images.

为了从互联网上挖掘有用的国防信息,课题组构建了面向国防科技领域的技术和术语语料库。该语料库中定义了基础技术、综合技术、武器、组织和军事术语 5 类技术术语。基于此标注语料,本文分别探索了子词单元和语言学信息在传统序列标注 Bi-LSTM+CRF 模型上的应用,并与最新基于

到稿日期:2019-03-16 返修日期:2019-07-30 本文受国家自然基金项目重点项目(61836007),面上项目(61772354,61773276)资助。
冯鸾鸾(1995-),女,硕士生,CCF 学生会员,主要研究方向为自然语言处理;**李军辉**(1983-),男,副教授,硕士生导师,主要研究方向为机器翻译、自然语言处理,E-mail:jhli@suda.edu.cn(通信作者);**李培峰**(1971-),男,教授,博士生导师,主要研究方向为自然语言处理和机器学习;**朱巧明**(1963-),男,教授,博士生导师,主要研究方向为自然语言处理。

BERT 模型的识别方法进行了对比。实验表明,子词信息和语言学信息的使用都能够有效地提升技术和术语的识别性能。同时,本文还在通用领域公开数据集 CoNLL03[1] 上进行了实验,进一步验证了本文方法的有效性。

2 相关工作

在研究领域,目前并未发现专门针对技术和术语识别的研究,但可以将技术和术语看作特定领域的"命名实体",采用命名实体识别类似方法进行技术和术语识别。命名实体识别就是将文本信息中规定的命名实体(即本文中的技术和术语)识别出来,其在自然语言处理中是一项基础性的工作,在信息抽取、机器翻译、自动问答等领域有着广泛的应用。

自从 1995 年在 MUC-6 评测会议上提出了命名实体的概念后[2],命名实体识别开始受到国内外研究者的广泛关注。所使用的方法可以粗略地分为三大类:基于规则的方法、基于统计机器学习的方法和基于深度学习的方法。

基于规则的方法[3]需要人工构建大量的规则集合,规则集合通常需要相关领域的专家参与构建,这种人工构建规则的方式依赖于具体文本风格,一般在小数据集上可以达到很高的性能,但是随着标注数据集的不断扩大,人工制定规则集合的方式成本高昂,并且某一领域上构建的规则集合很难移植到其他领域。因此,随着机器学习在自然语言处理领域的兴起,命名实体识别研究逐渐转向使用统计学知识和基于机器学习的方法,主要方法包括隐马尔可夫模型(Hidden Markov Models)[4-5]、最大熵模型(Maximum Entropy Models)[6-7]、支持向量机(Support Vector Machines)[8-9]和条件随机场(Conditional Random Fields)[10]等。基于统计机器学习方法的命名实体识别研究的一个重点是如何创建出有利于该任务的特征,包括词的上下文信息、词的位置、词之间的搭配、词本身的含义等信息。例如,McCallum 等[10]采用 CRF 方法进行命名实体识别,在 CoNLL-2003 英文数据集上,$F1$ 值达到了 84.04%。近年来,基于深度学习方法的命名实体识别也获得了广泛的成功。Huang 等[11]将双向循环神经网络 LSTM 和条件随机场 CRF 模型相结合进行命名实体识别,在 CoNLL-2003 英文数据集上 $F1$ 值达到了 90.10%。Ma 等[12]使用卷积神经网络 CNN 获取字符向量,并与词向量一起作为双向循环神经网络(Bi-LSTM)的输入,最后使用条件随机场 CRF 模型对输出的标签之间建立相互依赖关系,在 CoNLL-2003 英文数据集上 $F1$ 值达到了 91.21%。Bharadwaj 等[13]在 LSTM 神经网络上增加了一层音素特征,并利用 attention 机制在土耳其语等形态变化较复杂的语言上取得了较好的 NER 效果。Peters 等[14]提出 ELMo(Embeddings from Language Models)词表示,基于字符卷积的双层双向语言模型,在 CoNLL-2003 英文数据集上 $F1$ 值达到了92.22%。近年来,Devlin 等[15]提出了一种新型语言表示模型 BERT(Bidirectional Encoder Representations from Transformers),采用表义能力更强的双向 Transformer 网络结构来预训练语言模型,在 CoNLL-2003 英文数据集上 $F1$ 值达到了 92.8%,与 Akbik 等[16]提出的使用字符级语言模型为句子中的字符串生成动态的上下文嵌入方法的结果相当。

近年来,与国防科技领域相近的军事领域的命名实体识别研究越来越受到重视,Guo 等[17]基于规则和词典的方法,从战术报告中抽取有意义的实体。单赫源等[18]利用条件随机场模型学习文本特征,识别出作战文书中的部队、装备、地点和任务等命名实体。冯蕴天等[19]利用 CRF 模型提取相关军事文本特征,并结合军事词典及规则校正 CRF 模型的识别结果,能够较好地识别出军事文本中的人员军职军衔名、军事装备名、军用物资名、军事设施名、军事机构名(含部队番号)以及军用地名等军事命名实体。王学锋等[20]基于深度学习框架 Bi-LSTM+CRF,引入了字符向量表示,识别出了军事想定语料集中的部队、机构、时间、地名、武器、设施、环境和数量 8 类军事命名实体。

基于以上研究,本文探索了子词信息和语言学信息在技术和术语识别中的应用。具体地,基于 Bi-LSTM+CRF 深度学习模型,先使用 Bi-LSTM 获取字符向量或子词向量表示,然后与词向量、句法等语言学特征向量一起作为 Bi-LSTM 的输入,最后用 CRF 模型对输出序列进行优化,从而得到一个最优的识别结果,并与基于 BERT 模型的技术术语识别方法进行对比。

3 基于 Bi-LSTM+CRF 模型的技术和术语识别

3.1 基准模型:基于词的 Bi-LSTM+CRF 模型

本文使用 Bi-LSTM+CRF 作为实验的基准模型,该模型能自动学习词序列特征,如图 1 所示,其主要包含 3 个模块:词表示层、特征抽取层和序列标注层。

1)词表示层:对文本进行基础特征表示。神经网络无法直接处理自然语言,因此本文使用词向量[21]表示词汇信息。具体地,本文采用 glove 预训练的词向量[1)]对词汇进行初始化,在训练过程中对词向量进行更新。

2)特征抽取层:对文本表示进行特征抽取。本文使用 Bi-LSTM 通过训练进行词序列特征抽取,Bi-LSTM 既可以捕获上文信息,也可以捕获下文信息,即同时高效获得文本序列的上下文信息。

3)序列标注层:对经特征抽取层得到的上下文信息进行标注,采用 CRF 模型,通过在相邻标签间添加转移分数来获得标签依存关系,优化输出序列,从而得到最优的识别结果。

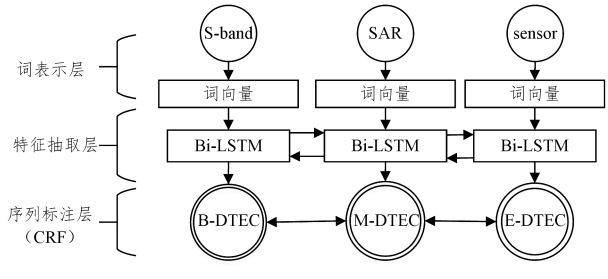


图 1 基于词的 Bi-LSTM+CRF 模型

Fig. 1 Word-based Bi-LSTM+CRF model

1) http://nlp.stanford.edu/projects/glove/

3.2　基于字符、子词的模型

Lample 等[22]利用 Bi-LSTM 在训练时抽取字符特征以挖掘单词之间的共性。图 2 给出了从构成单词的字符生成该单词的词的表示过程，即词的表示包含字符表示和词向量两部分，其中字符向量均为随机初始化[23]。

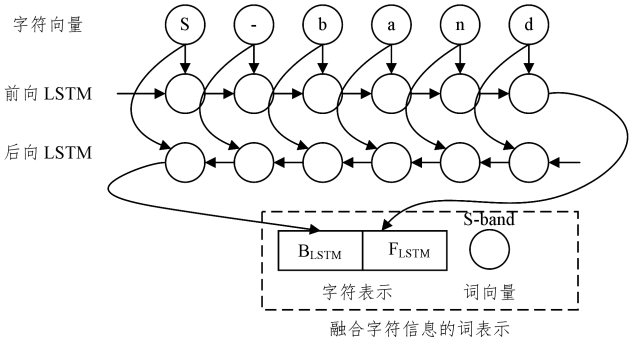


图 2　基于字符的 Bi-LSTM＋CRF 模型

Fig. 2　Char-based Bi-LSTM＋CRF model

另外，Sennrich 等[24]将子词应用到神经机器翻译以解决稀有词的翻译问题。子词既可以解决稀有词的表示问题，同时又较字符包含更多的信息量，如 anti-missile（反导导弹）、short-range（短程）、air-based（空基）等词常作为技术或术语的修饰词，而在通用领域，这些词通常被视为稀有词，本文将 anti-missile 分为 anti-和 missile 两个子词，anti-包含了“反”这一含义，将 short-range 分为 short 和-range 两个子词，分别表示“短”和“程”，将 air-based 分为 air-和 based 两个子词，air-表示“空中”。由此可见，子词对于复合词的拆分相比单个字符包含了更多的信息，考虑到面向国防科技领域的技术和术语大多包含复合词，本文应用基于字节对编码（BPE）压缩算法的分词技术，利用 Bi-LSTM 抽取子词特征，用于替代字符特征，子词特征比字符特征包含更多的信息，具体过程如图 3 所示。

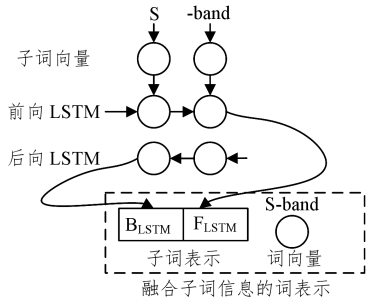


图 3　基于子词的 Bi-LSTM＋CRF 模型

Fig. 3　Subword-based Bi-LSTM＋CRF model

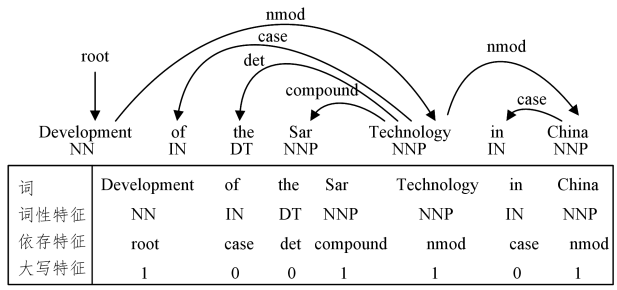
3.3　融入语言学特征的技术和术语识别

Sennrich 等[25]在神经机器翻译任务中将词性、依存句法等作为输入特征，提高了英德等翻译系统的性能。甘丽新等[26]在中文实体关系抽取中加入了句法语义特征。本文将探索适用于技术术语识别任务的语言学特征。

不同于传统意义的实体，面向国防领域的技术和术语通常具备如下特点：1）技术大多为复合名词短语，常含专有名词，且复合名词短语中常用名词或形容词进行修饰；2）武器类别因为是特定型号或名称，大多为首字母大写或全部大写，且名词居多；3）军事术语类别大多为首字母大写。例如，例 2 中，HJ-1-C 是 China Radar Earth-observation satellite 的一个命名，属于武器类别，拼字特征上字母全部大写；China Radar Earth-observation satellite 是一个技术，其中 China 和 Radar 词性识别为专有名词，修饰 Earth-observation satellite，并与其共同组成复合名词短语。

例 2：HJ-1-C is a China Radar Earth-observation satellite in a polar orbit at an altitude of 500 km，and has been launched on November 19，2012.

结合上述语料特点，本文选取词性、依存句法和大写 3 个语言学特征。其中，单词的依存句法特征为该单词与其中心词的依存关系；单词的大写特征分为 3 种情况，分别是全字母小写、首字母大写和全字母大写，分别用数字 0，1，2 表示。图 4 给出了句子“Development of the Sar Technology in China”的依存句法树以及 3 个语言学特征示例。



词	Development	of	the	Sar	Technology	in	China
词性特征	NN	IN	DT	NNP	NNP	IN	NNP
依存特征	root	case	det	compound	nmod	case	nmod
大写特征	1	0	0	1	1	0	1

图 4　依存句法树与语言学特征示例

Fig. 4　Example of dependency tree and linguistics features

为每个单词获取如上所述的 3 个特征后，本文将语言学特征作为输入特征，与字符表示或子词表示、词向量串联共同组成词表示，如图 5 所示。其中，“Sar”的字符或子词经过 Bi-LSTM 得到其相应的字符或子词表示，与“Sar”的词向量和其 3 个特征向量串联，组成“Sar”新的词表示，进入图 1 所示的后续 Bi-LSTM＋CRF 模型。

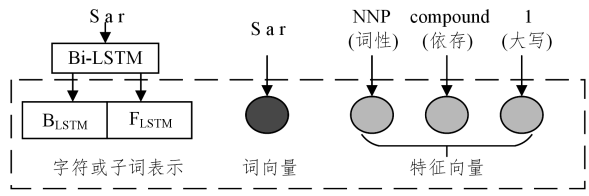


图 5　融合字符或子词信息和语言学信息的词表示

Fig. 5　Word representation of fusing character or subword information and linguistic information

4　实验

4.1　语料及实验设置

本文使用的面向国防科技领域的技术和术语语料库共包含 479 篇文章，总计 24 487 句，33 756 个技术术语。各类技术术语的统计如表 1 所列。本文采用 BMOES 标记格式，即 B 表示技术术语起始位置，M 表示中间位置，E 表示结束位置，S

单独构成技术术语，O 表示其他。针对 479 篇文章，按照 5∶1∶1 的比例随机选择 344 篇作为训练集，68 篇作为开发集，67 篇作为测试集。首先使用 NLTK 对语料进行分句处理，然后使用斯坦福大学(Stanford)的 CoreNLP[1] 工具对文档进行词条化(tokenize)、词性标注和依存句法分析。为获取子词，采用 BPE 算法进行子词化处理[2]，操作数设为 3000。

表 1 语料标注数据统计

Table 1 Annotation data statistics of ONDS corpus

类别	个数(比例/%)	类别	个数(比例/%)
基础技术	802(2.4)	综合技术	24808(73.5)
武器	3786(11.2)	组织	2138(6.3)
军事术语	2222(6.6)		

本文实验的参数设置如表 2 所列。采用的字符向量维度为 30 维，字符 LSTM 的隐层大小为 50 维，子词向量维度为 50 维，子词 LSTM 的隐层大小为 100 维，采用训练好的子词向量和 glove 预训练的词向量共同构成词表对子词进行初始化。每个特征向量维度为 10 维，词向量维度为 100 维，词 LSTM 的隐层大小为 300 维。使用随机梯度下降(SGD)算法训练模型，设置一个批次的样本数为 10，迭代次数为 100，学习率为 0.005，并采用 Hinton 等提出的 dropout 方法将隐层的节点以 0.5 的概率随机忽略[27]。基于 BERT 模型的技术术语识别方法均采用文献[15]中的默认参数。

表 2 实验参数

Table 2 Experiment parameters

参数	训练值	参数	训练值
字符向量维度	30	每个特征向量维度	10
字符 LSTM 隐层大小	50	迭代次数	100
子词向量维度	50	批样本数	10
子词 LSTM 隐层大小	100	学习率	0.005
词向量维度	100	dropout rate	0.5
词 LSTM 隐层大小	300		

实验评估标准采用准确率(Precision，P)、召回率(Recall，R)和 F1 值，计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{1}$$

$$R=\frac{TP}{TP+FN} \tag{2}$$

$$F1=\frac{2PR}{P+R} \tag{3}$$

其中，TP 表示识别出的技术术语中正确的数量；FP 表示识别出的技术术语中不正确的数量；FN 表示没有识别出的技术术语数量。

4.2 实验结果

表 3 列出了各模型的实验结果。其中，WLSTM-CRF 是仅基于词的 Bi-LSTM+CRF 模型，是本文方法的基准实验；CLSTM-WLSTM-CRF 是基于字符的 Bi-LSTM+CRF 模型；SLSTM-WLSTM-CRF 是基于子词的 Bi-LSTM+CRF 模型。

表 3 技术和术语识别的实验结果

Table 3 Experiment results of technology and terminology identification

(单位：%)

模型		P	R	$F1$
不融入语言学特征	WLSTM-CRF(基准模型)	71.76	66.01	68.76
	CLSTM-WLSTM-CRF	73.30	67.72	70.40
	SLSTM-WLSTM-CRF	72.66	68.52	70.53
融入语言学特征	WLSTM-CRF	70.80	69.35	70.07
	CLSTM-WLSTM-CRF	**74.18**	68.63	71.29
	SLSTM-WLSTM-CRF	73.45	70.22	**71.80**
BERT	BERT-Softmax	68.86	71.68	70.24
	BERT-CRF	69.98	72.63	71.28
	BERT-BiLSTM-CRF	69.58	**73.56**	71.51

表 3 同时列出了各模型在融入语言学特征前后的性能对比。在与基于 BERT 模型的对比实验中，本文复现了 3 个基于 BERT 模型的技术术语识别方案，分别是 BERT-Softmax 模型、BERT-CRF 模型和 BERT-BiLSTM-CRF 模型。

对比各个模型的实验结果可知，在融入了语言学特征后，基于子词的模型的实验性能最好，$F1$ 值达到了 71.80%，较基准系统提高了 3.04%。

4.3 实验分析

本文从 4 个方面对实验结果展开分析。

(1)各个模型的对比分析

从表 3 的实验结果可以看出：

1)基于子词的模型的实验性能最高。不融入语言学特征时，基于子词的模型的 $F1$ 值比基准模型的 $F1$ 值提高了 1.77%，基于字符的模型 $F1$ 值比基准模型的 $F1$ 值提高了 1.64%，基于子词的模型识别效果与基于字符的模型识别效果相当。融入语言学特征后，基于子词的模型 $F1$ 值比基准模型 $F1$ 值提高了 1.73%，基于字符的模型 $F1$ 值比基准模型 $F1$ 值提高了 1.22%，基于子词的模型 $F1$ 值比基于字符的模型 $F1$ 值提高了 0.51%。这说明字符序列和子词序列通过 Bi-LSTM 获得了一些仅用词向量训练抽取不到的上下文信息，并且子词比字符在一定程度上蕴含了更多的信息。

2)加入子词序列信息有助于提高召回率，其召回率比基准模型提高了 2.51%，说明子词序列信息有助于从文本中找出更多的技术术语。

3)加入字符序列信息有助于提高准确率，其准确率比基准模型提高了 1.54%，说明字符序列信息能够排除识别的部分假技术术语。

4)BERT 模型极大地提高了召回率，但准确率却下降了很多，由此可见，BERT 模型确实可以学到很多信息，有助于标注更多的技术术语，但同时提高了技术术语的错误率。

(2)语言学特征分析

从表 3 的实验结果可以看出：

1)在基准模型中融入语言学特征，与基于字符的模型识别效果相当，说明本文所提出的语言学特征在一定程度上可

[1] https://stanfordnlp.github.io/CoreNLP

[2] http://www.github.com/rsennrich/subword-nmt

以弥补字符序列通过 Bi-LSTM 获得的上下文信息。

2)在基准模型中融入语言学特征，$F1$ 值提高了 1.31%；在基于子词的模型中融入语言学特征，$F1$ 值提高了 1.27%；在基于字符的模型中融入语言学特征，$F1$ 值仅提高了 0.89%。语言学特征在没有字符序列信息的情况下可以有效提高实验性能，说明字符序列通过 Bi-LSTM 模型获得了一定程度的语言学特征信息。

3)融入语言学特征有助于提高召回率，在基准模型上提高了 3.34%，在基于子词的模型上提高了 1.7%，在基于字符的模型上提高了 0.91%。这说明语言学特征有助于从文本中找出更多的技术术语。

(3)各类别识别结果的分析

表 4 列出了基于子词模型的各类别的识别性能，可以看到，基础技术和军事术语的识别性能较低，综合技术、武器和组织的识别效果较好。一方面，综合技术在标注语料中数量最多，并且此类技术术语在句子中的语言学特征明显，武器和组织类别虽然数量不多，但这两类技术术语在句子中均有明显的语言学特征，例如组织类别大多为专有名词，并且首字母大写居多。另一方面，基础技术数量最少，并且大多为普通物理技术，例如 electrical energy(电能)，此类技术术语在句子中没有明显的语言学特征，军事术语与武器、组织类别对于模型而言有一定的困惑度，例如 Battle of the Beams(Blitz)是一个战争的名称，Blitz 意为闪电战，均应标注成军事术语类别，而模型将 Blitz 识别为武器，因为 Blitz 也可以作为某种武器的型号或名称。

表 4　各类别的识别结果

Table 4　Identification results of each category

(单位：%)

类别	P	R	$F1$
基础技术	44.58	37.00	40.44
综合技术	73.43	73.19	73.31
武器	74.69	75.22	74.96
组织	79.85	71.57	75.49
军事术语	73.80	39.43	51.40

(4)识别边界分析

由于面向国防科技领域的技术和术语大多由多个词语组成，计算完全匹配的正确率较为严格，因此本文计算了上述模型实验结果的左、右边界识别效果，如表 5 所列。

表 5　技术和术语左、右边界的识别效果

Table 5　Boundary identification results of technology and terminology

(单位：%)

模型		左边界 $F1$	右边界 $F1$
不融入语言学特征	WLSTM-CRF(基准模型)	72.62	74.32
	CLSTM-WLSTM-CRF	73.95	75.42
	SLSTM-WLSTM-CRF	74.47	75.71
融入语言学特征	WLSTM-CRF	73.65	75.17
	CLSTM-WLSTM-CRF	74.76	76.14
	SLSTM-WLSTM-CRF	75.56	76.41
BERT	BERT-Softmax	75.72	75.74
	BERT-CRF	75.96	76.12
	BERT-BiLSTM-CRF	76.07	76.33

在 4 个模型中，右边界的识别效果均在不同程度上好于左边界，说明对于本文的语料和识别任务而言，右边界更易被识别。融入语言学特征均有助于左、右边界的识别效果的提升，子词信息和 BERT 模型更有助于左边界的识别效果的提升，其中 BERT 模型的提升效果更加明显。

4.4　公开数据集实验结果

CoNLL03[1]是目前通用领域命名实体识别最常用的公开数据集。为了验证本文所提出的子词信息和语言学特征的有效性，在此数据集上进行相关实验，其在测试集上的结果如表 6 所列。

表 6　CoNLL03 数据集实验结果

Table 6　Experiment results of CoNLL03 dataset

(单位：%)

模型		P	R	$F1$
不融入语言学特征	WLSTM-CRF(基准模型)	90.91	88.35	89.61
	CLSTM-WLSTM-CRF	91.17	91.08	91.12
	SLSTM-WLSTM-CRF	89.85	88.24	89.04
融入语言学特征	WLSTM-CRF	91.21	90.72	90.96
	CLSTM-WLSTM-CRF	**91.23**	91.22	91.23
	SLSTM-WLSTM-CRF	91.15	**91.50**	**91.32**

由表 6 可以看出，本文所提出的语言学特征也适用于 CoNLL03 数据集，融入语言学特征的 3 个模型均比之前有了不同程度的提升。相比技术术语识别，子词模型识别通用领域命名实体的效果并不显著。分析语料，其原因如下：1)首先子词主要是为了解决稀有词的表示问题，而通用领域的语料稀有词数量较少；2)在本文所提出的技术术语语料中，复合词数量相比通用语料更多，子词对于复合词的拆分比单个字符包含了更多的信息。而在通用语料中应用子词化技术并不能达到预期的效果，反而有可能破坏原本的单词结构，导致实验效果下降。

结束语　针对面向国防科技领域的技术和术语识别，本文探索了子词单元在 Bi-LSTM+CRF 模型上的应用，采用深度学习与传统语言学特征相结合的方法进行技术和术语识别的实验，并与基于 BERT 模型的技术术语实验结果进行对比。实验结果表明，本文提出的方法的 $F1$ 值最高达到了 71.80%，能有效识别国防科技领域军事文本中的技术和术语。同时，在公开数据集 CoNLL03 上的实验结果也进一步验证了本文方法的有效性。

目前的实体识别方法都是以句子为单位，忽略了篇章信息。而对于本文的技术和术语识别任务，由于技术和术语通常具有较强的专业性，仅根据句子内容很难判断某个名词短语是否为技术或术语。通过实验结果进一步发现，同一篇章内的某个名词短语存在标注不一致的现象。因此，在未来的工作，将引入篇章信息，使模型在标注时能够参考篇章信息，同时对同一个名词短语增强其识别一致性，从而进一步提升识别效果。

参 考 文 献

[1] SANG K T, MEULDER D F. Introduction to the conll-2003 shared task: Language-independent named entity recognition [C]// Proceedings of the 2003 Conference on Natural Language Learning. 2003:142-147.

[2] CHINCHOR N. MUC-6 named entity task definition (version2.1) [C] // Proceedings of the 6th Conference on Message Understanding. Columbia, Maryland, 1995.

[3] COLLINS M, SINGER Y. Unsupervised models for named entity classification[C] // Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999: 100-110.

[4] ZHOU G D, SU J. Named Entity Recognition using an HMM-based Chunk Tagger [C] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. ACL, 2002: 473-480.

[5] BURGER J D, HENDERSON J C, MORGAN W T. Statistical named entity recognizer adaptation[C] // Proceedings of the 6th Conference on Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2002: 1-4.

[6] CHIEU H T, NG H T. Named Entity Recognition with a Maximum Entropy Approach[C] // Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. 2003: 160-163.

[7] CURRAN J R, CLARK S. Language independent NER using a maximum entropy tagger[C] // Proceedings of the Conference on Natural Language Learning at HLT-NAACL. 2003: 164-167.

[8] EKBAL A, BANDYOPADHYAY S. Named entity recognition using support vector machine: A language independent approach [J]. International Journal of Electrical and Electronics Engineering, 2010, 4(2): 155-170.

[9] MAYFIELD J, MCNAMEE P, PIATKO C. Named entity recognition using hundreds of thousands of features[C] // Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL. Stroudsburg: Association for Computational Linguistics, 2003: 184-187.

[10] MCCALLUM A, LI W. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons [C] // Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg: Association for Computational Linguistics, 2003: 188-191.

[11] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[EB/OL]. [2015-08-09]. https://arxiv.org/pdf/1508.01991.pdf.

[12] MA X Z, HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. ACL, 2016: 1064-1074.

[13] BHARADWAJ A, MORTENSEN D, DYER C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 1462-1472.

[14] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C] // Proceedings of NAACL-HLT 2018. New Orleans: Association for Computational Linguistics, 2018: 2227-2237.

[15] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. [2019-05-24]. https://arxiv.org/pdf/1810.04805.pdf.

[16] AKBIK A, BLYTHE D, VOLLGRAF R. Contextual String Embeddings for Sequence Labeling[C] // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 1638-1649.

[17] GUO J K, BRACKLE D V, LOFASO N, et al. Extracting meaningful entities from human-generated tactical reports[J]. Procedia Computer Science, 2015, 6(1): 72-79.

[18] SHAN H Y, ZHANG H S, WU Z L. A Military Named Entity Recognition Method Based on CRFs with Small Granularity Strategy[J]. Journal of Academy of Armored Force Engineering, 2017, 31(1): 87-88. (in Chinese)
单赫源，张海粟，吴照林. 小粒度策略下基于 CRFs 的军事命名实体识别方法[J]. 装甲兵工程学院学报，2017，31(1)：87-88.

[19] FENG Y T, ZHANG H J, HAO W N. Named Entity Recognition for Military Text[J]. Computer Science, 2015, 42(7): 15-18, 47. (in Chinese)
冯蕴天，张宏军，郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学，2015，42(7)：15-18，47.

[20] WANG X F, YANG R P, ZHU W. Military Named Entity Recognition Method Based on Deep Learning[J]. Journal of Academy of Armored Force Engineering, 2018, 32(4): 94-98. (in Chinese)
王学锋，杨若鹏，朱巍. 基于深度学习的军事命名实体识别方法[J]. 装甲兵工程学院学报，2018，32(4)：94-98.

[21] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C] // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013: 746-751.

[22] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C] // Proceedings of NAACL-HLT. San Diego, California, 2016: 260-270.

[23] YANG J, LIANG S L, ZHANG Y. Design challenges and misconceptions in neural sequence labeling[C] // Proceedings of the 27th International Conference on Computational Linguistics (COLING). 2018.

[24] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016.

[25] SENNRICH R, HADDOW B. Linguistic Input Features Improve Neural Machine Translation[EB/OL]. (2016-06-27). https://arxiv.org/pdf/1606.02892.pdf.

[26] GAN L X, WAN C X, LIU D X, et al. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features [J]. Journal of Computer Research and Development, 2016, 53(2): 284-302. (in Chinese)
甘丽新，万常选，刘德喜，等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展，2016，53(2)：284-302.

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