基于多标签的军事领域命名实体识别

单义栋 王衡军 王 娜
(中国人民解放军战略支援部队信息工程大学三院 郑州 450001)

摘 要 为了识别军事文本中的军事命名实体,根据军事命名实体的特点,将其分为 6 类标注。在此基础上,为了进一步解决多嵌套和组合的复合军事命名实体难以识别的问题,对传统的标注方法加以改进,提出了一种基于多标签的标注方法。首先,对复合的军事命名实体做分词处理,使之成为多个最小词组的组合;然后,各部分词组按其在命名实体中的位置做分段标注,各词组中的每个字则在分段标注的基础上,根据其在词组中的位置再做词位标注;最后,将整个标注作为军事命名实体中每个字的标注结果。实验结果表明,该标注方法能够提升军事命名实体的识别效果。

关键词 军事命名实体,多标签,复合军事命名实体
中图法分类号 TP391 **文献标识码** A

Military Domain Named Entity Recognition Based on Multi-label

SHAN Yi-dong WANG Heng-jun WANG Na
(The Third Institute, Information Engineering University, Zhengzhou 450001, China)

Abstract In order to identify military named entities in military texts, this paper classified them into six categories according to the characteristics of military named entities. On this basis, in order to further solve the problem that the multi-nested and combined composite military named entities are difficult to identify, the traditional annotation method was improved, and a multi-label annotation method was proposed. First, the compound military named entity is divided into several words, so that it becomes a combination of multiple minimum phrases, and then each part of the phrase is segmented according to its position in the named entity. On the basis of segmentation, each word in each phrase is marked with a vocabulary based on its position in the phrase. Finally, the entire label is ultimately used as the labeling result for each word in the military named entity. The experimental results show that the annotation method can enhance the recognition effect of military named entities.

Keywords Military named entity, Multi-label, Composite military named entity

1 引言

随着信息技术的发展,大量的信息系统被应用于各级军事指挥机构,使部队信息化作战能力得到显著提升。然而,信息化作战也使得需要处理的信息量大增,指挥员在面对大量的文电信息时如何快速准确地提取出所需信息,是当前亟待解决的关键问题。

军事命名实体是军事文本中各种命名实体的统称,主要包括军事地名、军事机构名、武器装备等。军事命名实体识别是军事信息处理的基础性工作,因此本文结合信息化条件下的作战特点来研究军事领域的命名实体识别问题。本文以军事需求为牵引,通过引入自然语言处理领域的命名实体识别技术对非结构化的军事文本信息进行分析与处理,从中提取出对作战行动与军事部署有用的关键信息,用以辅助指挥信息系统提升决策能力。

目前,军事领域的命名实体识别一般采用基于统计模型的方法如条件随机场(Conditional Random Field,CRF),或基于深度学习的方法如长短时记忆网络(Long-Short Term Memory network,LSTM)等。文献[1]结合军事领域术语的特点,选取词本身、词性、词长、是否在军事词典、前后窗口是否有特征符号以及互信息 MI 作为 CRF 模型的特征,实验结果显示该方法的 F 值达到了 72.05%。文献[2]利用 CRF 模型可以提取相关军事特征的优点,结合军事词典及规则,校正 CRF 模型的识别结果,达到了较好的军事命名实体识别效果。文献[3-4]利用 word2vec 工具训练词向量,计算词向量的相似度,并以此作为 CRF 模型的特征,实验结果表明,词向量的相似度特征能够较大幅度地提升模型性能。文献[5]提取含分词标签的字特征,并结合词性标注、关键词权值、依存句法 3 项词语特征共同作为 CRF 模型的特征,实验结果证明该方案能够提升命名实体的识别效果。与 CRF 模型相比,基于深度学习的方法可以自动选择和提取特征,不需人工选择,通常采用词向量[6]或字向量[7]作为输入,结合深度学习模型来实现命名实体的识别。文献[8]运用领域相关度与一致度相结合的特征词筛选算法,放大领域词与通用词之间的差异,压缩军事领域特征词数量,优化军事领域词向量,将准确率和召回率分别提升了 20%和 16.7%。文献[9]结合字向量与词向量,引入字词向量的概念,将字向量与其对应的词向量拼接作为输入,取得了优于单独以字或词作为输入的命名实体的识别效果。

然而,无论是添加各种 CRF 模型的特征与规则,还是结

单义栋(1988-),男,硕士,主要研究方向为自然语言处理,E-mail:1405478343@qq.com;**王衡军**(1973-),男,博士,副教授,主要研究方向为神经网络、机器学习,E-mail:2793730105@qq.com(通信作者)。

合深度学习的字词向量的标注方式，都只对简单的命名实体识别效果较好，而无法充分表达复合命名实体的内部结构关系，因此对复合命名实体的识别效果不尽如人意。

为提高军事领域命名实体尤其是军事文本中大量复合军事命名实体的识别效果，高强等[10]采用层叠模型，在底层利用规则识别简单的命名实体，而后将识别的结果传递给高层的 CRF 模型，辅助高层识别复合的军事命名实体。与单层模型相比，该模型能够有效识别部分复合的军事机构名，但存在严重依赖军事词典的规模以及人工选取规则不完善的问题。文献[11-12]也利用了层叠模型，与高强等人不同的是，在高层利用支持向量机(Support Vector Machine，SVM)和 CRF 模型相结合的方法，通过 SVM 识别左右边界词，而后将左右边界词作为 CRF 模型的特征，以达到提高复合军事组织机构名识别效果的目的。但 SVM 模型是通过特征词识别边界词，特征词表的不完备会导致部分复合军事实体不能准确识别的问题，因此需要不断完善特征词表，工作量较大。

针对复合军事命名实体识别率不高的现状以及结合 CRF 模型对军事领域命名实体识别存在人工提取军事特征不足的缺陷，本文提出一种基于深度学习的多标签军事命名实体标注方式，重点提升复合军事命名实体的识别效果。将复合军事命名实体拆分为多个不可再分词组，词组按其在命名实体中的位置做分段标注，词组内部单个字再根据其在词组中的位置做词位标注，最后将所有标注结合作为该字最终的标注结果。实验结果表明，该标注方法可以有效提高军事命名实体的识别效果。

2 军事领域命名实体的标注及识别模型

2.1 军事命名实体的分类

通用的命名实体识别一般包括人名、地名、组织机构名 3 类，因此本文也将人名、军事地名、军事组织机构名 3 类列为需识别的军事命名实体。与通用文本不同的是，军事文本中通常还包含有军事领域独特的命名实体，主要包括以下 3 类。

(1)军事职务职级：与其他领域相比，部队存在严格的上下级关系，这种关系通常以职务、职级来体现，职务名常以“长”字为右边界，如“连长”“营长”“团长”等，职级由“正”或“副”字与“军”“师”“团”等代表部队层级的字符组成，如“正连”“副团”等。这类军事实体名的组成方式固定，且特征明显。

(2)军事武器装备：军事文本中通常包含大量的武器装备名称，它们通常由字母、数字代表的装备型号和汉字代表的装备名组成，如“T-55 坦克”“M1 式加兰德步枪”。

(3)部队编制：该类实体通常以“军”“师”“团”等为边界字符，且通常以组合的方式出现，如“XX 师 XX 团”。

因此，本文将军事命名实体定义为 6 类：人名、军事职务职级、军事地名、军事机构名、军事武器装备、部队编制。

2.2 军事命名实体的标注方式

军事命名实体分为简单军事命名实体和复合军事命名实体。简单军事命名实体一般构成简单，识别相对容易，如连长、正连、彭德怀等。而复合军事命名实体一般由简单命名实体和相关词汇组成，如黄埔军校、济南军区 XX 军 XX 师等，识别过程中很容易将一个复合的军事命名实体识别为多个与原文语义不符的词组。这类军事命名实体结构复杂，通常内部又嵌套简单的命名实体，识别困难。

在军事命名实体的标注过程中，简单的军事命名实体采用通用的字向量标注方式，而对复合军事命名实体而言，通用的标注方法除了实体的第一个字标注为实体首部外，其他所有字标注相同，无法准确区分，导致许多有价值的信息被忽略。本文将其切分成多个具有实际意义的词组，这些词组不可再分，如“黄埔军校”可分为“黄埔”和“军校”两个词组，而后词组中每个字根据词组在命名实体中的位置以及该字在词组中的位置，标注为一个多层标签，最终根据标签结果，将其重新组合成一个完整的军事命名实体。这种多标签的标注方式可以根据军事命名实体内部不同部分不同字的位置做不同的标注，因此可以得到更多反映内部结构的特征，有助于提升军事命名实体的识别效果。

2.3 军事命名实体标签

本文以字为单位标注，简单的军事命名实体采用命名实体标签集与分段标签的联合标签标注。标签集包括人名、军事职务职级、军事地名、军事机构名、军事武器装备、部队编制。分段标签集采用 BIOS 四标签，其中 B 代表实体开始、I 代表实体内部、S 代表单字实体、O 代表非实体。

针对嵌套、组合的复合军事命名实体，采用实体标签集、分段标签集、词位标签集联合标注。实体标签集包括军事地名、军事机构名、军事武器装备、部队编制；分段标签集与简单军事命名实体的分段标签集相同；词位标签集采用 BMES 四标签，分别代表各词组的开始词位、中间词位、结束词位及单字词组。简单军事命名实体和复合军事命名实体的标注如表 1 和表 2 所列。

表 1 简单军事命名实体标注

序号	命名实体	实体首部	实体内部及结尾
1	人名	PER-B	PER-I
2	军事职务职级	ZW-B	ZW-I
3	简单军事地名	LOC-B	LOC-I
4	简单军事机构名	ORG-B	ORG-I
5	简单武器装备	WEP-B	WEP-I
6	简单部队编制	MIL-B	MIL-I

表 2 复合军事命名实体标注

序号	命名实体	实体首部	实体内部及结尾
1	复合军事地名	LOC-B-B	LOC-I-B
		LOC-B-M	LOC-I-M
		LOC-B-E	LOC-I-E
		LOC-B-S	LOC-I-S
2	复合军事机构名	ORG-B-B	ORG-I-B
		ORG-B-M	ORG-I-M
		ORG-B-E	ORG-I-E
		ORG-B-S	ORG-I-S
3	复合武器装备	WEP-B-B	WEP-I-B
		WEP-B-M	WEP-I-M
		WEP-B-E	WEP-I-E
		WEP-B-S	WEP-I-S
4	复合部队编制	MIL-B-B	MIL-I-B
		MIL-B-M	MIL-I-M
		MIL-B-E	MIL-I-E
		MIL-B-S	MIL-I-S

例如：句子“济南军区陆军某旅开展比武竞赛”，其标注序列为[MIL-B-B，MIL-B-M，MIL-B-M，MIL-B-E，MIL-I-B，MIL-I-E，MIL-I-B，MIL-I-E，O，O，O，O，O，O]。

2.4 军事命名实体识别模型

BLSTM 可以自动学习到上下文的特征而不需要人工选

择，但最终的输出却是单个输入最优结果的拼接而不是整个序列的最优结果；CRF 考虑的却是优化整个序列，对各状态之间的依赖关系建模。两者各有优劣，却又可以相互弥补，因此本文利用 BLSTM-CRF 模型来达到更好的效果。BLSTM-CRF 模型如图 1 所示。

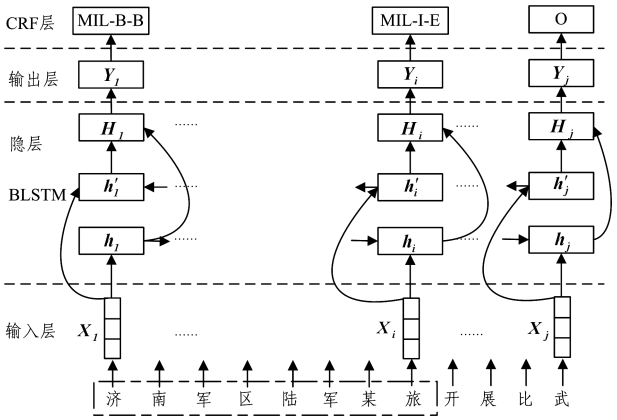

图 1　BLSTM-CRF 模型

2.4.1　输入层

计算机无法识别军事文本中的汉字，因此首先需要将文本中的汉字向量化。向量化有两种方式：独热向量表示和分布式向量表示。独热向量虽然简单，但存在严重的数据稀疏问题，而且无法表达字与字之间的语义信息。分布式向量以低维度向量表示，且相关字之间的向量表示更加接近，因此在自然语言处理任务中一般采用分布式向量表示。

文献[5]的实验结果表明，预训练的字向量可以提升命名实体的识别效果，因此本文利用谷歌的 word2vec[13] 工具将语料库中的每一个汉字转化为固定长度的分布式向量，所有的向量生成字向量表，表中不同字对应不同的向量，同一个字对应唯一的向量。通过查表可以查到输入序列中的每一个汉字对应的向量，并以此作为命名实体识别模型的输入。

2.4.2　隐层

长短时记忆网络目前被广泛应用于自然语言理解任务中，与循环神经网络(Recurrent Neural Network，RNN)相比，新增了记忆单元和门机制，以解决长距离信息依赖问题。记忆单元用于储存信息，而这些信息受 3 个门机制的控制，分别为遗忘门、输入门、输出门。遗忘门用于去除记忆单元中的部分旧信息，输入门用于决定哪些新信息需要存入记忆单元，输出门决定输出记忆单元的哪些信息。

单向 LSTM 只能保存前文信息，无法有效利用后文信息，因此具有一定的局限性。为充分运用前文和后文的信息，本文在隐层采用双向长短时记忆网络(Bi-directional Long-Short Term Memory network，BLSTM)。BLSTM 包含前向和后向两个 LSTM 网络，可以分别从前端和后端开始运行，每个输出单元连接前向和后向的两个 LSTM 单元，因此可以存储两个方向的信息，提升了模型性能。

将输入层字向量序列($\boldsymbol{X}_1$，…，$\boldsymbol{X}_i$)输入到隐层 BLSTM 中，相关公式表示如下：

$$\begin{cases}\boldsymbol{h}_i = f(\boldsymbol{U}\boldsymbol{X}_i + \boldsymbol{V}\boldsymbol{h}_{i-1}) \\ \boldsymbol{h}_i' = f(\boldsymbol{U}'\boldsymbol{X}_i + \boldsymbol{V}'\boldsymbol{h}_{i-1}') \\ \boldsymbol{H}_i = g(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{W}'\boldsymbol{h}_i')\end{cases} \tag{1}$$

其中，$\boldsymbol{X}_i$ 代表输入；$\boldsymbol{h}_i$ 表示前向 LSTM 的输出；$\boldsymbol{h}'$表示后向 LSTM 的输出；$\boldsymbol{U}$，$\boldsymbol{V}$，$\boldsymbol{U}'$，$\boldsymbol{V}'$，$\boldsymbol{W}$，$\boldsymbol{W}'$分别代表权重，初始计算时随机初始化，模型训练完成后得到各权重的值；$\boldsymbol{H}_i$ 表示将前向 LSTM 与后向 LSTM 在每个位置的隐状态相结合得到的输出结果。

2.4.3　输出层

本层将隐状态输出向量 $\boldsymbol{H}_i$ 的维度映射到 K 维(K 为标签类别的数量)，并通过 softmax 函数进行归一化处理，计算公式如式(2)所示。

$$\boldsymbol{Y}_i = \text{softmax}(\boldsymbol{\omega}\boldsymbol{H}_i + \boldsymbol{b}) \tag{2}$$

其中，$\boldsymbol{\omega}$ 代表权重，$\boldsymbol{b}$ 代表偏置项。

2.4.4　CRF 层

$\boldsymbol{Y}_i$ 是 $\boldsymbol{X}_i$ 所对应的各标签的概率矩阵，最终输出结果是取 $\boldsymbol{Y}_i$ 最大值所对应的标签。然而，选取的标签存在不符合约束规则的情况，例如标签 PER-B 后接标签 ORG-E。为保证整个序列的标注结果符合标签间的依赖关系，引入转移矩阵 $\boldsymbol{T}$，元素 T_{ij} 表示从标签 i 转移到标签 j 的概率。该层运用维特比算法，计算结果由 BLSTM 输出矩阵 $\boldsymbol{Y}$ 与转移矩阵 $\boldsymbol{T}$ 共同计算得到，整个序列的预测输出为：

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{T}_{l_{i-1}, l_i} + \sum_{i=1}^{n} \boldsymbol{Y}_{i, l_i} \tag{3}$$

2.4.5　训练

2.4.5.1　前向传播

在前向传播过程中，为防止神经网络模型发生过拟合现象，通常采用 dropout 技术。在模型训练过程中，dropout 使神经单元以一定的概率停止工作，不参与本次神经网络的训练与优化。研究表明[14-15]，该技术能够有效降低错误率，提高模型性能。

2.4.5.2　反向传播

神经网络训练的目标是使真实值与计算值的差值最小，即它们之间的损失函数最小。本文通过梯度下降法更新参数来优化模型，直至损失函数达到最小值。

(1)采用交叉熵作为损失函数，计算公式为：

$$E_k(Y_k, y_k) = -Y_k \log y_k \tag{4}$$

$$E(Y, y) = \sum_k E_k(Y_k, y_k) = -\sum_k Y_k \log y_k \tag{5}$$

其中，k 表示输出标签的类别数，$E_k(Y_k, y_k)$表示输出为第 k 类标签的损失函数，Y_k 表示输出为第 k 类标签的真实值，y_k 表示输出为第 k 类标签的预测值。

(2)求解损失函数对模型参数的偏导数

在计算模型参数梯度时，需要采用链式法则，通过中间变量求偏导得到梯度。以计算输出层的权重 $\boldsymbol{\omega}$ 为例，计算公式为：

$$\boldsymbol{o}_i = \boldsymbol{\omega}\boldsymbol{H}_i + \boldsymbol{b} \tag{6}$$

$$\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\omega}} = \sum_k \frac{\partial \boldsymbol{E}_k}{\partial \boldsymbol{\omega}} = \sum_k \frac{\partial \boldsymbol{E}_k}{\partial \boldsymbol{o}_k} \frac{\partial \boldsymbol{o}_k}{\partial \boldsymbol{\omega}} \tag{7}$$

(3)权重更新

$$\boldsymbol{\omega} = \boldsymbol{\omega} + \left(-\lambda \frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\omega}}\right) \tag{8}$$

其中，λ 为学习率。采用上述方法更新模型中所有的权重和偏置项，使损失函数达到最小值，完成模型的训练。

3　实验

3.1　实验环境

实验机的主要软硬件参数及神经网络参数设置如表 3 和表 4 所列。

表 3 实验环境配置表

类型	配置及版本
硬件	CRU 2.6GHZ Inter(R) Core(TM) i5-4300M，操作系统 Window 7，RAM 8GB
软件	Python 3.5.2，Anaconda3，Tensorflow0.12.0

表 4 参数设置表

参数	数值
字向量长度	64
隐层结点数	128
双向长短时记忆网络层数	2
dropout/%	20
学习率	0.0001
Batch_size	32
Epoch	10

3.2 实验数据集及评价指标

由于目前缺乏专业的军事领域语料库，因此本次实验所采用的语料库是经过人工标注生成的军事领域语料库，语料来源是通过 Scrapy 爬虫获取中国国防报 2018 年 1 月至 3 月的语料。共选取军事领域新闻 300 篇，其中训练集 210 篇、开发集 30 篇、测试集 60 篇，语料库标注分别采用传统的标注方式和本文所采用的多标签标注方式。

实验以准确率(P)、召回率(R)、综合指标值(F)作为评价标准，计算公式为：

$$P=\frac{\text{正确识别的实体数}}{\text{识别的实体数}}\times 100\% \tag{9}$$

$$R=\frac{\text{正确识别的实体数}}{\text{实体总数}}\times 100\% \tag{10}$$

$$F=\frac{2\times P\times R}{P+R}\times 100\% \tag{11}$$

3.3 实验结果及分析

采用 BLSTM-CRF 模型进行两组对比实验，其中实验 1 所用语料以通用的军事命名实体标签集与分段标签的联合标签标注，实验 2 所用语料以本文提出的多标签标注。两组实验的实验结果如表 5 所列

表 5 命名实体识别的对比结果

(单位：%)

实验模型	准确率	召回率	F 值
BLSTM-CRF 模型＋通用标签	89.07	85.91	87.46
BLSTM-CRF 模型＋多标签	92.28	88.63	90.42

由表 5 可知：实验 2 的结果优于实验 1，准确率、召回率、F 值分别提升 3.21%，2.72%，2.96%，说明本文提出的多标签标注方法在军事命名实体识别，尤其是在复合军事命名实体识别方面，比实验 1 所用的标注方法有更好的表现。分析原因，是本文所采用的方法使军事命名实体尤其是复合军事命名实体中每个字的标签更加精确，单字的区分度更高，在模型训练过程中会提取更多的特征，可以达到更好的识别效果。

本文在尽量还原前人实验环境的情况下，在相同的军事语料库下将实验结果与前人的工作进行对比。文献[4]构建双层模型，底层识别简单的军事命名实体，上层将 SVM 与 CRFs 相结合来辅助识别复杂的军事命名实体；文献[16]采用 CRF 模型识别小粒度的军事命名实体，再将小粒度组合成完整的军事命名实体。对比实验结果如表 6 所列，与文献[4]、文献[16]的实验结果相比，本文所提出的多标签军事命名实体识别方法的准确率、召回率、F 值都有提高，在 F 值上分别提高了 7.6%和 3.94%，达到了较高的识别水平。

表 6 与前人工作的对比

(单位：%)

实验模型	准确率	召回率	F 值
文献[4]	85.22	80.55	82.82
文献[16]	87.98	85.03	86.48
本文	92.28	88.63	90.42

结束语 针对军事命名实体识别的需求，本文根据军事文本中军事命名实体的特点，确定了 6 类需识别的军事命名实体。为提升军事命名实体识别效果，采用多标签的标注方法对复合的军事命名实体重新标注。实验结果表明：在军事语料的测试中，其准确率达到 92.28%，召回率达到 90.42%，取得了较好的识别效果。由于目前缺乏专业的军事语料库，需要人工获取和标注，工作量很大，制约了军事命名实体识别方面的研究。同时，复合军事命名实体在军事文本中所占比例相对较小，需要大量包含复合军事命名实体的训练语料训练模型，在一定程度上限制了多标签标注下军事命名实体效果的进一步提升。因此，在以后的工作中，如何在数据规模较小和标注少的情况下提升军事命名实体的识别效果，将是研究的重点。

参 考 文 献

[1] 田俊玮. 军事领域中文术语抽取的研究[D]. 大连：大连理工大学，2013.

[2] 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.

[3] 蒋超. 研报领域的产品词命名实体识别的研究[D]. 南宁：广西大学，2017.

[4] 姜文志，顾佼佼，胡文萱，等. 基于多模型结合的军事命名实体识别[J]. 兵工自动化，2011，30(10)：90-93.

[5] 孙安，于英香，罗永刚，等. 序列标注模型中的字粒度特征提取方案研究——以 CCKS2017：Task2 临床病历命名实体识别任务为例[J]. 图书情报工作，2018，62(11)：103-111.

[6] 章成志，苏新宁. 基于条件随机场的自动标引模型研究[J]. 中国图书馆学报，2008(5)：89-94，99.

[7] 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.

[8] 秦杰，曹雷，彭辉，等. 一种面向军事文本的领域特征词向量描述方法[J]. 计算机工程，2016，42(8)：160-165.

[9] 谢志宁. 中文命名实体识别算法研究[D]. 杭州：浙江大学，2017.

[10] 高强，游宏梁. 基于层叠模型的国防领域命名实体识别研究[J]. 现代图书情报技术，2012(11)：47-52.

[11] 乌兰敖日格乐. 中文军事组织机构名的识别[D]. 大连：大连理工大学，2010.

[12] 张磊. 特定领域命名实体识别通用方法的研究[D]. 北京：北京交通大学，2018.

[13] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济，2015，25(2)：145-148.

[14] 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.

[15] BOUTHILLIER X，KONDA K，VINCENT P，et al. Dropout as data augmentation[J]. arXiv：1508.08700.

[16] 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)：84-89.