文章编号：2095-2163(2019)04-0096-04 中图分类号：TP391 文献标志码：A

基于 BiLSTM_Att 的军事领域实体关系抽取研究

朱珊珊[1]，唐慧丰[2]
（1 信息工程大学洛阳校区，河南 洛阳 471003；2 信息工程大学，郑州 450001）

摘 要：军事领域中实体关系的抽取是该领域相关体系知识图谱建设的重要步骤。本文设计了基于 BiLSTM 和注意力模型（Attention）的实体抽取模型，该模型分为词向量表示、句子上下文特征提取以及关系分类三个阶段。在词向量表示阶段，模型创新性地加入词性特征。在对相关语料进行实验验证的基础上，结果显示该模型对军事类实体关系抽取有较好的 *F* 值。

关键词：关系抽取；BiLSTM_Att；向量表示；词性特征

Research on military domain entity relationship extraction based on BiLSTM_Att

ZHU Shanshan[1], TANG Huifeng[2]
(1 Luoyang Campus, Information Engineering University, Luoyang Henan 471003, China;
2 Information Engineering University, Zhengzhou 450001, China)

【Abstract】 The extraction of entity relations in the military field is an important step in the construction of knowledge maps of related systems in this field. In this paper, a physical extraction model based on BiLSTM and Attention is designed. The model is divided into three stages: word vector representation, sentence context feature extraction and relationship classification. In the stage of word vector representation, the model innovatively adds part-of-speech features. On the basis of experimental verification of relevant corpus, the results show that the model has a good *F* value for military entity relationship extraction.

【Key words】 relational extraction; BiLSTM_Att; vector representation; part of speech characteristics

0 引 言

作为国家政治集体的军事武装力量，军队有着严格的组织关系，且具有分工明确、又可以联合联动的关系特性。对于军事类实体进行关系抽取是丰富军队军事结构资料库，构成完整明晰关系网的重要组成部分。

近年来，FreeBase、DBpedia、百度百科等知识库的建设为诸多互联网应用提供了可靠的数据来源。知识图谱作为一种智能、高效的信息组织形式，能够将实体本身以及实体的各类关系以网状连接的图谱形式完整地描述出来，并进行可视化的展示，是一种清晰明了的数据内容及其内部关系展示形式。

知识图谱的发展经历了 3 个时代。知识图谱早期被称为本体时代。2001 年随着 Wikipedia 出现，知识图谱进入语义网时代。前期 2 个阶段的知识图谱构建方式包括人工编辑和自动抽取，但自动抽取方法主要是基于在线百科中结构化信息而忽略了非结构化文本，而互联网中大部分的信息恰恰是以非结构化的自由文本形式呈现。与链接数据发展的同期，许多知识获取的方法被提出，这些方法大多基于信息抽取技术，用以构建基于自由文本的开放域知识图谱。随着信息抽取技术的不断进步，2012 年 Google Knowledge graph 上线，自此进入了知识图谱时代。

早期的实体和关系抽取，包括实体关系的特征设计、语料的标注等，基本都是由人工完成的。但是由于自然语言处理的标注工具使用因人而异，并且人工选择的特征会直接影响到关系抽取和分类的效果，因此即使耗费巨大的人力物力，关系抽取的效果也并非十分理想。而基于深度学习的神经网络模型则可以通过多层次网络分析对大规模文本语料自动挖掘特征信息[1]。例如，循环神经网络在捕捉句子的上下文信息方面有着良好表现，可以反映一个句子中多实体间的关系。但循环神经网络对长距离依赖不够，因此本文使用双向长短时记忆网络（BiLSTM）捕获句子更多的上下文信息。同时，在对单词进行向量表示时，除了加入位置信息外，还加入词性特征，并使用注意力机制提取语句层面的特征，根据最后输出向量进行分类，完成实体关系抽取任

作者简介：朱珊珊（1995-），女，硕士研究生，主要研究方向：自然语言处理、数据挖掘；唐慧丰（1973-），男，博士（后），教授，主要研究方向：智能信息处理、机器学习。
通讯作者：唐慧丰 Email：peimingshanshan@163.com

收稿日期：2019-05-15

务。

1　相关研究

在知识图谱的发展需求推动下，关系抽取的方法从上世纪后半叶的基于人工编写规则的方法，逐渐发展到基于统计的方法，直至近十年来基于机器学习神经网络方法的陆续涌现[2]。

早期基于规则的方法虽然促进了关系抽取研究的长足进步，但其自身的局限性也很明显，如：人工编写规则的过程较复杂、规则产生的效率较低、系统针对性好、通用性差等，所以后来的研究逐渐又转向基于统计的方法。随着网络开放程度增加，以及电子元器件计算速度、存储能力的提升，文本数据体量和规模迅速增长。基于统计的方法开始快速发展并获得广泛应用，主要包括监督学习、Bootstrap 方法、远程监督学习、无监督学习等。

基于统计的学习方法，首先需要大量完整已进行实体标注和实体间关系标注的语料库，然后根据定义的关系类型和定义的实体类型，通过提取文本特征，将词特征、位置特征等通过不同的分类算法训练模型，在测试时根据训练的模型抽取训练语料的实体对，并判断其关系类型。由于在特征提取的过程中需要依赖自然语言处理的自动分词、词性标注等工具，就使得在对语料处理时工具操作中所造成正确率损失，会对最终的分类性能产生影响。除此之外，文本特征提取过程还需要参照专家经验，因此特征的设计和验证需要耗费大量人力物力。但统计方法不仅可以在无标注文本中抽取出实体对及其关系，也在一定程度上脱离了对领域知识的依赖。

近十年来，深度学习成为实体关系抽取中颇受业界瞩目的研究新方法，深度学习是一种特殊的机器学习方法，具有灵活性好、性能高等特点。相比于基于统计的方法，深度学习的神经网络模型可以自动获取文本特征，并不需要对文本特征进行复杂的设计和验证。基于深度学习神经网络模型的关系抽取方法和基于统计的监督方法相比主要有 2 个优势，可阐释分述如下。

(1)在字、词、短语等结构上统一使用低维、连续的向量表示，具体根据不同模型需要的不同颗粒度进行调整。

(2)在更大单元，即句子、篇章等向量表示上，使用不同的神经网络模型组合各类较小语言单元的特征向量。

研究中选用深度学习框架下的神经网络模型，对特征进行抽取和选择是自动完成的，因此其在效率和正确率上也超过了传统的基于统计的机器学习方法。

2　BiLSTM_Att 模型

为了表示更丰富的上下文信息，模型选取双向 LSTM，即 BiLSTM 对提取的词向量进行特征表示，随后加入注意力模型(Attention)对神经网络的输出进行加权，在此基础上输出关系分类的结果。因此 BiLSTM_Att 模型分为 3 个阶段，即：首先，进行词的向量表示；然后，是 BiLSTM 融合上下文信息；最后，是注意力模型对 LSTM 的输出训练权重矩阵。该模型的框架设计如图 1 所示。这里，拟对此展开研究论述如下。

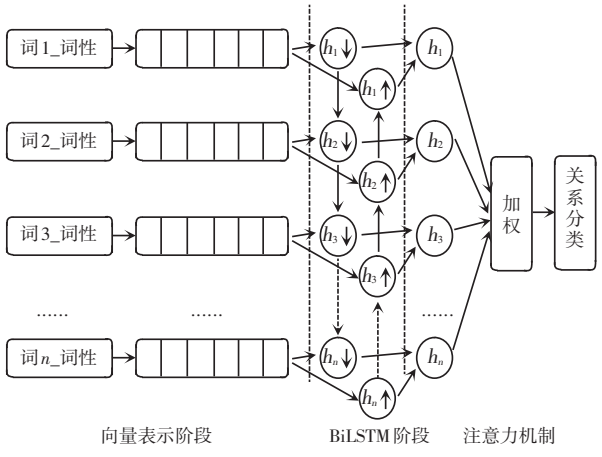

图 1　BiLSTM_Att 模型

Fig. 1　BiLSTM_Att model

2.1　加入词性的词向量表示

对词进行向量表示主要包括 2 个部分。一是词语本身的词向量训练，在训练过程中加入了词性信息。二是词的位置特征，指的是一个词距离该句子中 2 个实体词的位置关系。

在词向量训练前，根据词性标注结果，输入的词由“词-词性”表示，例如句子“<e1>Evo Morales<e1> has put <e2>Bolivia<e2> on the map.”经过预处理并加入词性信息后输入为“/Evo Morales_n /has_v /put_v /Bolivia_n /on_p /the_rzt /map_n”。由于 word2ver 是对 word embedding 的优化，因此本文的词向量训练使用 word2ver 工具中的 CBOW 模型。CBOW 模型的输入是一个词对应的上下文词的词向量，而输出是该词的词向量。例如一个句子片段“… distributed representations which encode the relevant grammatical relations…”上下文大小为 6，输出词是“encode”，那么输出的是“encode”的前 3 个

词和后 3 个词的词向量。需要说明的是,这 6 个词是没有先后顺序的,使用了词袋模型。该模型的训练过程中,研究定义了词向量的维度大小 M, 以及 CBOW 的上下文大小 $2c$, 这样对于训练样本中的每一个词,其前面的 c 个词和后面的 c 个词作为 CBOW 模型的输入,所有词汇词向量 $\boldsymbol{w}$ 作为输出。

除此之外,由于 word2ver 训练词向量使用的是词袋模型,没有包含词的位置信息,因此文本加入了词的位置向量以描述位置信息。例如在句子"<e1> Evo Morales<e1> has put <e2>Bolivia<e2> on the map."中,单词"has"距离"Evo Morales"和"Bolivia"两个实体分别为 1 和-2。将单词相对"head entity"和"tail entity"的距离映射成 2 个距离向量,组合词向量成为这个单词的向量表示。

该阶段对句子中词向量训练结束后,得到的是一个实数矩阵并传递给下一层,矩阵中包括了一个句子所有词的特征信息。

2.2 BiLSTM

LSTM 最早由 Hochreiter 和 Schmidhuber[3] 提出,为了解决循环神经网络中的梯度消失问题。主要思想是引入门机制,从而能够控制每一个 LSTM 单元保留的历史信息的程度以及记忆当前输入的信息,保留重要特征,丢弃不重要的特征。为了将上文信息和下文信息都进行表征,本文采用双向 LSTM,将上一个细胞状态同时引入到输入门、遗忘门以及新信息的计算当中。该 LSTM 模型也同样包含 4 个部分,如图 2 所示。由图 2 研究可知,其功能设计过程可解析概述如下。

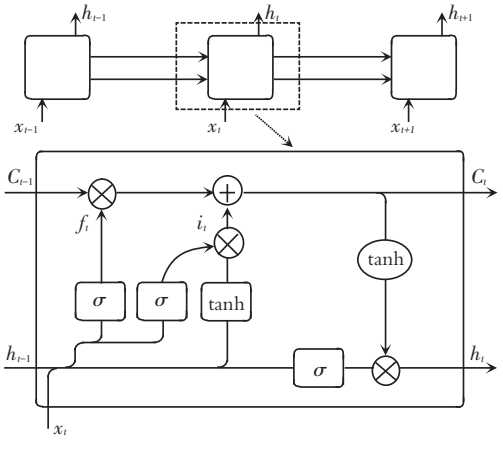

图 2 LSTM 模型

Fig. 2 LSTM model

输出门包含了当前输入、上一个隐状态、上一个细胞状态,组成权重矩阵,以决定加入多少新信息。对应的数学公式为:

$$i_t = \sigma(W_{xi}\, x_t + W_{hi}\, h_{t-1} + W_{ci}\, c_{t-1} + b_i), \quad (1)$$

遗忘门则决定丢弃多少旧的信息。对应的数学公式为:

$$f_t = \sigma(W_{xf}\, x_t + W_{hf}\, h_{t-1} + W_{cf}\, c_{t-1} + b_f), \quad (2)$$

细胞状态包含了上一个细胞状态以及基于当前输入和上个隐状态层信息生成的新信息。对应的数学公式为:

$$c_i = i_t\, g_t + f_t\, c_{t-1}, \quad (3)$$

$$g_t = \tanh(W_{xc}\, x_t + W_{hc}\, h_{t-1} + W_{cc}\, c_{t-1} + b_c), \quad (4)$$

输出门则包含了当前输入、上一个隐状态、当前细胞状态,组成权重矩阵,以决定哪些信息被输出。对应的数学公式为:

$$o_t = \sigma(W_{xo}\, x_t + W_{ho}\, h_{t-1} + W_{co}\, c_t + b_o), \quad (5)$$

最终,输出的当前隐状态可由当前细胞状态乘以输出门的权重矩阵得到。对应的数学公式为:

$$h_t = o_t \tanh(c_t). \quad (6)$$

2.3 Attention 机制

注意力模型是从心理学上的注意力模型中引入的。人脑的注意力模型指的是,当一个人看到了整幅画面时,在特定的时刻 t, 人的意识和注意力的焦点是集中在画面中的某一个部分上,其它部分虽然还在人的眼中,但是分配给这些部分的注意力资源是很少的。深度学习中的注意力机制从本质上看和人类观察事物的选择性视觉注意力机制类似,就是从视觉所观察范围内的众多信息中选择核心观察点,也就是对完成当前任务最重要的一部分信息。

在本文模型中,将 LSTM 层输入的向量集合表示为 H:$[h_1, h_2, \cdots, h_T]$。其 Attention 层得到的权重矩阵由下面的方式得到 :

$$\boldsymbol{M} = \tanh(H), \quad (7)$$

$$\alpha = Softmax(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}), \quad (8)$$

$$r = \boldsymbol{H}\,\boldsymbol{\alpha}^{\mathrm{T}}, \quad (9)$$

其中, $\boldsymbol{H} \in R^{dw\times \mathrm{T}}$;$d_w$ 为词向量的维度;$\boldsymbol{w}^{\mathrm{T}}$ 是一个训练学习得到的参数向量的转置。最终用以分类的句子将表示如下:

$$h^* = \tanh(r). \quad (10)$$

2.4 Softmax 分类器

在 Attention 模块后加入一个 Softmax 分类器,用来预测标签 $\hat{y}$。该分类器将上一层得到的隐状态作为输入。研究推得计算公式具体如下:

$$\hat{y}(y \mid S) = Softmax(W^{(s)}\ h^* + b^S), \quad (11)$$

$$\hat{y} = Softmax\ \hat{p}(y \mid S), \quad (12)$$

成本函数采用正样本的负对数的似然函数,研

究推得计算公式具体如下：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda \|\theta\|_F^2. \quad (13)$$

其中，t 为正样本的独热(one-hot)表示；$\hat{y}$ 为 Softmax 估计出的每个类别的概率；m 为类别个数；λ 是正则化的超参数。

3　实验验证及结果分析

相比于无领域关系抽取，军事类实体关系抽取要在更大程度上受制于军队组织机构隶属关系、人员隶属关系以及武器装备系统的分队等。因此，针对军事领域实体关系抽取，本文选取了 3 000 条相关语料进行标注，其中涉及到的实体关系共有 7 种，详见表 1。

表 1　实体关系类型

Tab. 1　Entity relationship type

序号	名称	释义
1	Unknown	不明
2	共事	2 个人之间为同事或战友关系
3	上下级	2 个人之间为上级对下级或下级对上级
4	校友	2 个人属于同一培养单位成员
5	隶属	人员隶属于单位，或武器装备隶属于单位
6	平行	隶属于同一上级的 2 个同级单位
7	归属	下级单位归属于上级单位，或船只飞机等归属某一舰队

对 3 000 条标注语料进行筛选，补全核对标注信息，并进行预处理后，将其中的 2 500 条作为训练语料，500 条作为测试语料。各个类别测试结果见表 2。

表 2　测试结果

Tab. 2　The test results

类别	准确率	召回率	F 值
1	0.718 3	0.787 2	0.752 8
2	0.842 9	0.769 9	0.806 4
3	0.722 0	0.701 4	0.711 7
4	0.882 4	0.790 3	0.836 4
5	0.795 4	0.813 6	0.804 5
6	0.811 7	0.759 1	0.785 4
7	0.862 4	0.829 3	0.845 9

测试结果显示，“校友”关系和“归属”关系的整体识别率较高，但是“上下级”关系的识别效果不理想，并且该关系类型也是召回率最低的。

4　结束语

文本使用 BiLSTM_Att 模型完成了对军事类中文语料的关系抽取任务。该模型由加入词性和位置信息的词向量训练、双向 LSTM 上下文特征抓取以及注意力模型的权重分配三个阶段组成。在对语料进行实验后发现，该模型整体效果较好，但是对于“上下级”、“平行”关系类型的识别召回率还是略有逊色。因此，在接下来的实验中，应更关注于实体关系抽取召回率的提升。除此之外，对军事领域关系抽取的语料建设也应有所关注。

参考文献

[1] 庄成龙，钱龙华，周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报，2009，23(1)：3-8，34.

[2] 车万翔，刘挺，李生. 实体关系自动抽取[J]. 中文信息学报，2005，19(2)：1-6.

[3] HOCHREITER S，SCHMIDHUBER J. Long short-term memory [J]. Neural Computation，1997，9(8)：1735-1780.

[4] RINK B，HARABAGIU S. Utd：Classifying semantic relations by combining lexical and semantic resources[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala，Sweden：Association for Computational Linguistics，2010：256-259.

[5] KAMBHATLA N. Combining lexical，syntactic，and semantic features with maximum entropy models for extracting relations [C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Barcelona，Spain：Association for Computational Linguistics，2004：1-5.

[6] 杜嘉，刘思含，李文浩，等. 基于深度学习的煤矿领域实体关系抽取研究[J]. 智能计算机与应用，2019，9(1)：114-118.

[7] 万静，李浩铭，严欢春，等. 基于循环卷积神经网络的实体关系抽取方法研究[J/OL]. 计算机应用研究：1-6[2018-12-26]. http://kns.cnki.net/kcms/detail/51.1196.TP.20181225.1615.003.html.