面向东北亚地区的军事舆情文本关系抽取方法

隗　昊[1,2]　金百川[1]
WEI Hao　JIN Baichuan

摘　要　随着信息技术的迅速发展，军事舆情监测在地区安全与国际关系中扮演着愈发重要的角色。东北亚地区作为全球政治、经济和安全格局的重要支点，面向该地区构建军事舆情知识图谱能够有效地分析和观测舆情热点，对于协助政府决策、构建和谐稳定的国际关系具有重大价值。关系抽取是构建知识图谱的关键技术和核心任务，受到研究者们广泛的关注。以新闻网站中与东北亚地区相关的军事舆情文本作为数据源，建立包含 7 种实体类型、11 种关系类型的东北亚军事舆情实体关系数据集，并提出基于 BERT-CapsNet 架构的舆情关系抽取模型。实验结果表明，所提出的方法在所构建的舆情实体关系数据集上具有较高的准确率和召回率，能够有效地识别出文本中的实体关系，为后续的军事舆情监测提供了有力的数据支持和技术支撑。

关键词　东北亚地区；军事舆情；关系抽取；深度学习；预训练语言模型

doi：10.3969/j.issn.1672-9528.2024.06.003

0 引言

随着信息技术的快速发展，军事舆情在地区安全与国际关系中的影响日益凸显。军事舆情文本作为反映地区军事动态、揭示潜在冲突风险的重要信息源，对于各国决策层而言具有极高的战略价值。东北亚地区作为世界政治、经济和安全格局的重要支点，对其军事舆情的分析与解读尤为重要[1]。因此，研究面向东北亚地区的军事舆情文本关系抽取方法，对于深入了解地区军事动态、预测潜在冲突风险、推断事态演化动向具有重要意义[2]。

关系抽取是指从文本中自动抽取出实体间的关系，进而揭示文本中实体的深层关联信息和潜在交互模式[3]。特定领域文本的特殊性，如标注训练语料的稀缺性、语言表述的复杂性、信息分布的碎片化以及领域知识的专业性等，使得东北亚军事舆情文本关系抽取成为一项具有挑战性的任务。传统的文本关系抽取方法往往难以有效处理这些问题，因此需要探索新的方法和技术，以适应军事舆情文本关系抽取的需求。本文首先针对目前东北亚地区军事舆情标注语料匮乏问题，收集了新闻网站中与东北亚相关的军事新闻，并将这些数据进行清洗和整理，提出了面向东北亚地区的军事舆情实体关系体系，建立东北亚军事舆情实体关系数据集。然后提出一种基于深度学习的军事舆情文本关系抽取模型，将预训练语言模型和胶囊网络进行结合，对军事舆情文本进行语义建模和特征提取，实现实体间关系的自动抽取。最后模型通过在本文所构建的数据集上进行实验，验证了所提方法的有效性。

1 相关研究

关系抽取对知识图谱的构建起到十分重要的作用，是自然语言处理领域中的一个重要的子任务，其目的是从非结构化文本中识别和提取实体之间的关系。早期研究人员主要依赖手工设计的规则和模板实现关系抽取。随着机器学习的兴起，监督学习方法逐渐在实体关系抽取中占据主导地位，研究人员开始使用标注好的训练数据，构建机器学习模型来学习从文本中提取实体和关系的模式。近年来，由于深度学习技术的兴起，卷积神经网络[4]和长短期记忆网络[5]等神经网络模型被提出，以及大规模预训练模型的引入（如 BERT[6]、GPT[7] 等），使得基于深度学习的关系抽取方式成为主流。在特定领域关系抽取研究中，王欢等人[8]使用预训练模型 FinBERT 提出一种面向金融文本的关系抽取方法。党小超等人[9]提出一种融合多种神经网络的 Voting 模型，对矿井提升系统故障知识图谱构建的关系抽取环节有了显著的提升。张鲁等人[10]提出一种基于图结构的实体关系模型 RoGCN-ATT，为地质知识图谱的构建和应用提供了有力的支持。在军事舆情关系抽取领域，尤其是中文实体关系抽取

1. 大连外国语大学软件学院 辽宁大连 116044
2. 大连外国语大学中国东北亚语言研究中心 辽宁大连 116044
[基金项目] 辽宁省高等学校基本科研项目“基于事理图谱的东北亚多语言文本舆情监测方法研究”（LJKQZ20222451）

方面，目前仍然处于起步阶段。当前存在的挑战之一是相关数据集和模型相对有限，因此在如何有效地将深度学习等先进技术应用到军事舆情实体关系抽取中，仍需要进一步深入研究和探索。

2 基于 BERT-CapNet 的关系抽取模型

2.1 模型框架

本文以 bert-base-chinese 中文预训练语言模型进行嵌入表示，并结合胶囊网络 [11] 进行序列编码，提出了一种 BERT-CapsNet 模型。这种组合尝试充分利用 BERT 在自然语言处理任务中的预训练表示学习能力，同时结合胶囊网络的层次化特征学习和动态路由机制，以更好地捕捉输入数据中的层次性结构和关系。BERT-CapNet 网络结构如图 1 所示。

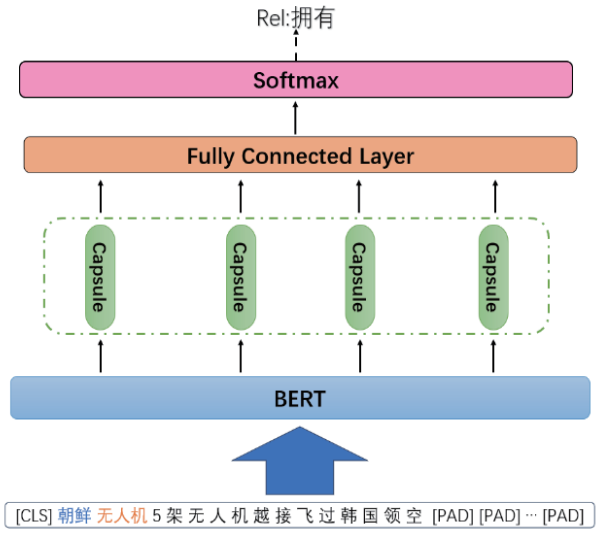

图 1 BERT-CapNet 模型结构图

2.2 BERT 预训练语言模型

BERT 是一种基于 Transformer 架构的深度学习模型，专门设计用于自然语言处理任务，其核心思想在于利用双向上下文信息来预训练深层的语言表示。如图 2 所示，BERT 的核心结构包括以下几个关键部分。

（1）多层 Transformer 编码器：BERT 模型由多个 Transformer 编码器组成，这些编码器分为多层。每一层都由多头自注意力机制和前馈神经网络组成。这种层次结构允许 BERT 在不同层次上学习不同粒度的语言表示。

（2）输入嵌入：BERT 使用三种嵌入表示，分别为标记嵌入、段嵌入和位置嵌入，这些嵌入的组合形成了模型的输入表示，使其能够理解输入文本的结构和语境。

（3）双向上下文建模：与传统的语言模型不同，BERT 采用了双向的上下文建模，在预训练过程中，模型同时考虑了输入序列中每个位置左右两侧的上下文信息，从而更全面地捕捉输入样本的语境信息。

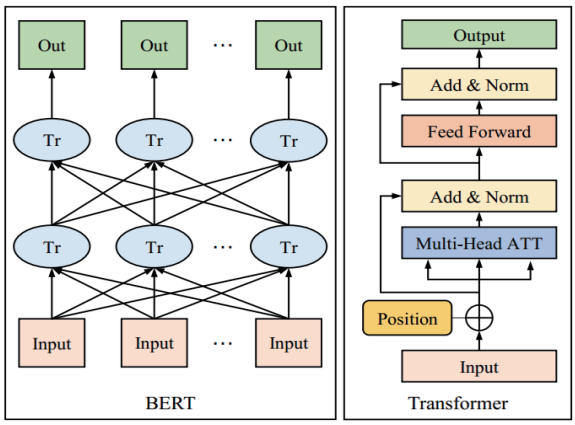

图 2 BERT 预训练语言模型结构图

2.3 胶囊网络

胶囊网络由 Geoffrey Hinton 等人于 2017 年提出，旨在克服传统卷积神经网络中存在的一些问题，如姿态变化、空间层次关系不明确等。它引入了胶囊（Capsule）的概念，以更有效地学习和表示层次结构化的特征。胶囊网络应用在自然语言领域可以提取文本中更深层次的特征。胶囊网络中的“胶囊”是网络中的神经元，与传统神经元中每个神经元输出的一个标量不同，胶囊的输出是一个向量，这个向量的长度和方向编码了实体的各种属性，向量的长度表示实体存在的概率。胶囊网络主要通过动态路算法和 squash 激活函数进行前向传播。动态路由算法流程图如图 3 所示，是胶囊网络中用于决定如何将信息从一层胶囊传递到下一层胶囊的机制，其目的是使网络能够学习实体部分与整体之间的层次关系。动态路由的基本步骤为：

$$s_j = \sum_i c_{ij} \hat{\mu}_{j|i} \tag{1}$$

$$\hat{\mu}_{j|i} = W_j \mu_i \tag{2}$$

$$c_{ij} = soft\max\left(b_{ij}\right) \tag{3}$$

$$squash\left(x\right) = \frac{\| x \|^2}{1+\| x \|^2} \frac{x}{\| x \|} \tag{4}$$

$$v_j = squash\left(s_j\right) \tag{5}$$

$$b_{ij} = b_{ij} + \hat{\mu}_{j|i} v_j \tag{6}$$

式中：s_j 为模型输入，$\hat{\mu}_{j|i}$ 为浅层胶囊输出，W 为参数矩阵，squash 为压缩函数，经过非线性变换后得到深层胶囊 v_j，整个动态路由过程迭代进行，以生成最终的特征输出。

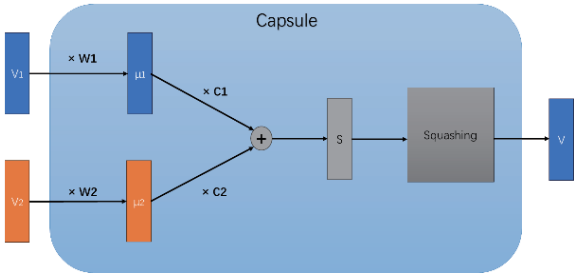

图 3 动态路由算法

3 实验与分析

3.1 数据来源与处理

目前没有公开的东北亚军事舆情领域文本数据集，因此本文收集了 2022 年至 2023 年新浪网东北亚国家军事主题新闻语料。对所有收集到的数据进行了清洗和整理，构建了东北亚军事舆情数据集，数据集的构建主要包括实体及关系的定义和语料标注等步骤。首先，本文构建了如图 4 所示的包含 7 类实体、11 类关系的体系结构，实体类型分别为：“人物”“国家 _ 国家地区”“机构”“部队”“武器装备”“军事活动”“政策”。人物实体是指涉及与军事相关的个体，如领导人、指挥官等。国家 _ 国家地区实体是指涉及国家或国家某个地区，可包括世界各国及地区。机构实体包含了大学、研究所等研究机构，以及国家的外交行政等政府机构。部队实体涉及各种军事力量，如陆军、海军、空军、特种部队以及相关军事单位。武器装备实体涉及军事用途的武器和装备，包括坦克、飞机、舰船、火箭炮等军事装备。军事活动实体涉及军事行动及时间，包括军演、战役、军事训练等。政策实体涉及各国政府制定的相关政策，包括国防政策、外交政策、战略计划等。实体之间的关系分为 11 种，分别为：“反对 _ 竞争”关系表示两个国家之间存在冲突或者在外交方面存在谴责；“支持 _ 合作”关系表示两个国家存在外交、战略支持以及合作研究等；“拥有”关系表示一个国家拥有或者制造某种武器装备；“地点”关系表示某项军事活动在某个国家或者某个国家具体的地区所发生；“使用 _ 装备”关系表示部队使用配备某种武器装备；“制定”关系是指国家制定了某项政策；“生产”关系表示某个机构生产某种武器装备；“参与”关系是指某个部队参与了特定的军事活动；“成员”关系是指某个人物是某个机构的成员；“国籍”关系是指某个人物的国籍是某个国家；“开展”关系是指某一个国家指挥开展了某种军事活动。

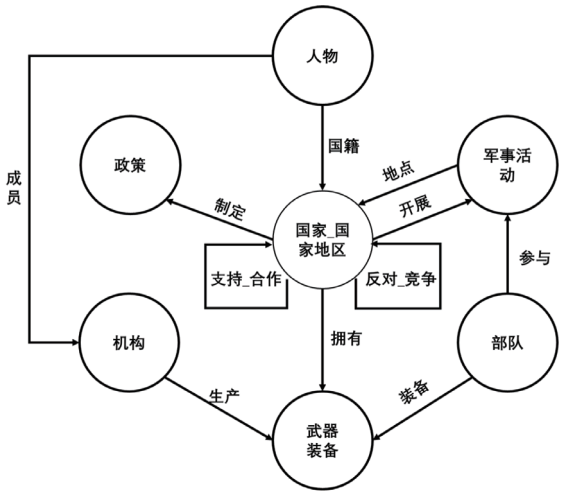

图 4 东北亚军事舆情实体关系体系

在完成实体关系体系的设计后，对所收集的东北亚军事舆情语料进行标注，最终构建的数据集语料总量为 3030 条，其中“拥有”关系数据有 705 条，“地点”关系数据有 60 条，“反对 _ 竞争”关系数据有 389 条，“使用 _ 装备”关系数据有 120 条，“制定”关系数据有 139 条，“支持 _ 合作”关系数据有 419 条，“生产”关系数据有 22 条，“参与”关系数据有 54 条，“开展”关系数据有 182 条，“成员”关系数据有 185 条，“国籍”关系数据有 755 条。

3.2 评估指标

为了评估 BERT-CapNet 模型对东北亚军事舆情文本的关系抽取效果，本文采用精确率（precision，P）、召回率（recall，R）和 F_1 值（F_1-score）作为性能评价指标，具体计算方法为：

$$P=\frac{\text{预测正确的样本数}}{\text{预测出的所有样本数}}\times 100\% \tag{7}$$

$$R=\frac{\text{预测正确的样本数}}{\text{所有真正的样本数}}\times 100\% \tag{8}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{9}$$

式中：精确率 P 衡量了模型预测的所有正例样本中实际为正例的比例，召回率 R 衡量了模型能够正确检测出的正例样本的比例，F_1 值是精确率和召回率的调和平均值。

3.3 与现有模型的性能比较

为了验证本文模型在东北亚军事舆情数据集上的有效性，采用了多个基线模型进行多次对比实验，并对各个模型实验结果进行汇总和比较分析，实验结果如表 1 所示。

表 1 模型性能对比

模型	P	R	F_1
BERT-LSTM	95.54	95.39	95.47
RoBERTa	96.08	96.01	96.05
BERT	95.78	95.72	95.75
BERT-CapNet	96.24	96.22	96.23

由表 1 结果分析可知，BERT 采用了 Transformer 结构，其准确率 P、召回率 R、F_1 值分别为 0.957 8、0.957 2 和 0.957 5。BERT-LSTM 模型在 BERT 的基础上引入了 LSTM 网络，但在性能上略微降低，准确率、召回率和 F_1 值为 0.955 4、0.953 9 和 0.954 7，表明在这个特定任务中，引入 LSTM 并没有显著提升模型性能，可能因为 BERT 本身已经具备了对序列信息的较好捕捉能力，过于复杂的模型可能会导致过拟合现象的出现。RoBERTa 模型通过对 BERT 的改进，表现出更高的性能水平，准确率、召回率和 F_1 值达到了 0.960 8、0.960 1 和 0.960 5，这进一步验证了对 BERT 进行改进的有效性，特别

是在训练步骤和批量大小等方面的优化，为模型性能的提升提供了有力支持。BERT-CapNet 模型引入胶囊网络结构，取得了最佳性能，准确率、召回率和 F_1 值分别为 0.962 4、0.962 2 和 0.962 3，这表明胶囊网络在这个任务中发挥了积极作用，可能通过其分层特征表示的方式更有效地学习了任务相关的信息，使得性能相较于其他模型有了较好的提升。

3.4 模型的训练损失监测

各模型的训练损失曲线监测结果如图 5 所示。由训练损失曲线可知，对于 BERT 模型，前 10 个 epoch 的损失从 0.004 2 迅速下降至 0.000 77，表现出较快的拟合速度；在 10 至 20 个 epoch 间，损失波动较小，可能已经相对稳定。RoBERTa 模型同样在前 10 个 epoch 内表现出较快的损失下降，从 0.052 0 降至 0.000 80，而后损失波动在一个较低水平，显示了相对平稳的拟合趋势。相反，BERT-LSTM 模型的损失在前 10 个 epoch 内从 0.088 6 下降至 0.008 83，下降速度较慢，并在接下来的 epoch 间波动，表现出较为平缓的拟合趋势。最后，BERT-CapNet 模型在前 10 个 epoch 内损失从 0.086 5 迅速下降至 0.002 14，显示出相对较快的拟合速度，而后在 10 至 20 个 epoch 间损失波动较小，表现出相对平稳的拟合趋势。

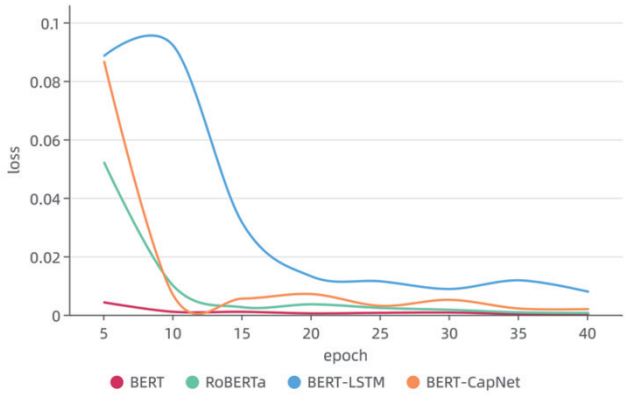

图 5 模型的训练损失曲线

4 结论

本文围绕东北亚军事舆情关系抽取任务进行研究，提出了基于 BERT-CapNet 的关系抽取模型，同时构建了一个东北亚军事舆情实体关系数据集，并在该数据集上进行了各项对比实验。实验结果表明，本文提出的 BERT-CapNet 模型相较于其他主流的基线模型取得了较好的性能表现。同时，本文所构建的数据集也为东北亚军事舆情实体关系抽取研究提供了更为有效的数据基础。在未来的研究中，将继续探索更加有效的方法，以提高东北亚军事舆情关系抽取的准确性，同时也将不断扩充数据集，为以后东北亚军事舆情领域的研究提供更可靠的技术支撑。

参考文献：

[1] 张蕴岭，杨伯江，项昊宇．百年变局纵深演进下东北亚局势回顾与展望 [J]. 东北亚学刊，2024(2):1-13+143.

[2] 王玥，赵健，朱燕．知识图谱军事运用的前景展望 [J]. 信息系统工程，2023(9):24-27.

[3] 李冬梅，张扬，李东远，等．实体关系抽取方法研究综述 [J]. 计算机研究与发展，2020,57(7):1424-1448.

[4]BERTONI F, CITTI G, SARTI A. LGN-CNN:a biologically inspired CNN architecture[J].Neural networks,2022,145:42-55.

[5] 陈前华，胡嘉杰，江吉，等．采用长短期记忆网络的深度学习方法进行网页正文提取 [J]. 计算机应用，2021, 41(S1): 20-24.

[6]DEVLIN J, CHANG M W, LEE K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,vol.6.long and short papers.Stroudsburg,PA:Association for Computational Linguistics, 2018: 4171-4186.

[7] 王浩畅，刘如意．基于预训练模型的关系抽取研究综述 [J]. 计算机与现代化，2023(1):49-57+94.

[8] 王欢，王兴芬，吕金娜．面向金融文本的实体关系抽取方法 [J]. 计算机工程与设计，2023,44(11):3345-3351.

[9] 党小超，叶汉鑫，董晓辉，等．矿井提升系统的故障实体关系抽取研究 [J/OL]. 计算机工程与应用 :1-11 [2024-03-10]. http://kns.cnki.net/kcms/detail/11.2127.TP.20230828.0904.002.html.

[10] 张鲁，段友祥，刘娟，等．基于 RoBERTa 和加权图卷积网络的中文地质实体关系抽取 [J/OL]. 计算机科学 :1-11[2024-03-05].http://kns.cnki.net/kcms/detail/50.1075.TP.20231129.1535.002.html.

[11] 杨巨成，韩书杰，毛磊，等．胶囊网络模型综述 [J]. 山东大学学报 (工学版), 2019, 49(6):1-10.

【作者简介】

隗昊（1993—），通信作者（weihao1005@163.com），男，山东济南人，博士，讲师，研究方向：自然语言处理、数字人文。

（收稿日期：2024-04-11）