基于多特征融合的文本关系抽取研究

解文康 [1,2]
XIE Wenkang

摘 要 数字化时代，互联网数据增长态势迅猛。在此背景下，如何从海量文本中精准、高效地抽取关系，并进一步挖掘出更具价值的信息，已成为自然语言处理、数据挖掘等相关领域的热门研究方向。因此，研究关系抽取方法具有重要的理论和实践价值。然而，现有的关系抽取方法大多存在两个主要限制：一是数据集的泛化性有限，二是主流方法在任务间的信息特征交互能力有限。基于此，文章提出了一种创新方法，采用变换器网络进行多特征融合，并利用共享组件来更好地适配实体对齐和解码任务，从而构建一种基于多特征融合的文本关系抽取模型，旨在提升关系抽取的效果，为验证模型有效性，分别在 Duie 和 NYT 数据集上进行了多组对照实验，均取得了最优效果，验证了本模型中共享组件模块与采用变换器网络结构进行多特征融合可以提升模型在中英文数据集上抽取的性能与泛化性。

关键词 深度学习；信息抽取；关系抽取

doi：10.3969/j.issn.1672-9528.2025.02.016

0 引言

近年来，互联网技术催生了数据的大量增长，不断迭代更新。海量的数据不断充实人类的知识库，使其日益丰富多元，为社会发展和知识创新提供了强大的数据支撑。与此同时，随着数据量的不断攀升，对大规模文本数据进行合理处理的需求也愈发迫切。因此，如何准确且有效地抽取实体之间的关系并挖掘出更有价值的信息，成为亟待攻克的关键课题。在这一背景下，关系抽取技术已经成为广大学者探索关注的前沿领域。

关系抽取是信息抽取中的一个重要任务，关系抽取技术是一种从纯文本中抽取关系事实的方法，通过识别文本上下文中的实体并对提及的这些实体之间存在的关系进行分类。关系抽取旨在提取出自然语言中实体对及其之间的语义关系 [1-2]。然而，现有主流研究方法均是针对英文关系抽取进行实验，如何让模型学习更多关系，灵活地推广到中文文本上成为当务之急。由于中英文数据集在语言表达、句子结构上存在较大差距，同时中英文数据中存在数量众多的多义词，并且中英文数据的表达方式与组织方式也大相径庭，因此提高模型在中英文数据抽取的适配性成为一个主流问题。使得对于关系抽取的研究仍存在巨大的挑战。

早期的关系抽取研究一般采取的是基于机器学习的方法，但其存在大量的局限性，比如难以处理大规模数据，以及对于特征工程的过度依赖。

随着研究的不断深入，基于深度学习的联合关系抽取方法逐渐流行。深度学习的关系抽取方法大体上可以分为流水线方法和联合抽取方法 [3]。然而，流水线方法的抽取存在误差传播问题，如果实体识别出现错误，会直接导致关系判断识别错误，从而得出错误的关系实体对三元组。联合抽取方法的关系抽取模型则克服了这个缺点 [4-5]。通过共享联合模型的编码层，实体识别和关系抽取子任务可以进行联合学习，实现两个子任务之间的相互依赖。这种方法直接输出文本中包含的三元组，避免了传统流水线方法中的错误累积问题。因此，现有主流研究方法都是基于实体和关系的联合学习。例如，文献 [6] 中通过一种新的视角设计了联合关系抽取框架，采用特殊的层叠式指针标注组件，将关系抽取任务分解成二重提取框架。这种方法首先使用主体标注器指针式标记主体，然后利用主体捕捉相应的客体与关系。文献 [7] 提出了一种基于端对端的信息抽取统一框架，简单的指令调优就可以适配不同的信息抽取任务。文献 [8] 提出了单步走的思想，有效解决了之前联合抽取分步走中的级联误差和冗余信息问题，设计了一个特定的角标分类器和相应的评分分类器，以评估三元组并进行关系抽取，但以上方法存在任务之间联系不足和在不同数据集上的泛化性差等问题。本文提出了一种融合

1. 三峡大学湖北省水电智能视觉监测重点实验室 湖北宜昌 443002
2. 三峡大学计算机与信息学院 湖北宜昌 443002

多特征的语义增强级联框架模型（FRGC），旨在解决主流模型中的关键问题。FRGC 模型在特征提取阶段采用 RetNet 网络 [9]，融合了关系类型特征，从而提高了模型在中英文关系抽取任务中的泛化性。同时，模型设计了一个特征共享交互模块，实现了实体对齐和序列标注模块之间的资源共享，提升了实体与关系匹配的准确度。实验结果表明，在 NYT 数据集 [10] 和 Duie 数据集 [11] 上，FRGC 模型表现出显著的有效性。

1 语义增强级联框架模型

文献 [12] 提出了一种基于潜在关系和全局对应组件的联合关系三重提取框架，该框架通过设计的潜在关系组件，将实体预测限制在潜在关系判断的子集里，同时利用全局通信组件来解码出对应三元组，从而降低复杂度。本文在此基础上进行了改进，不仅保留了 PRGC 的优势，还增强了词向量表征，融合了实体类型特征与关系特征到词向量中，以获取实体之间的潜在联系。具体来说，采用了 RetNet 网络层与多头自注意力机制结合的结构来强化获取的表征信息。使实体抽取模块以及全局实体信息标记模块能够充分融入强化的表征信息，从而提升关系映射客体的准确性。另外，本模型设计了一个通信共享组件，该模块使得全局对齐模块与序列标注模块可以进行信息共享，以此来提升实体对齐以及标注的准确性。本节主要介绍改进后的模型 FRGC，其整体架构如图 1 所示。

1.1 FRGC 编码器模块

本文中 FRGC 编码器主要由词向量编码器（sectence encoder）和特征强化编码器（featrue enhancement）组成。对于关系抽取任务，高质量的词向量表征信息会帮助模型对于关系抽取实体抽取的判断，那么如何得到一个优质的表征信息以及采用何种方法去丰富表征信息成为了一个难点，本文通过采用词向量编码器与特征强化编码增强器融合实体的类型特征以及实体对关系特征信息，以此挖掘实体的潜在联系，增强词向量与其他特征之间的信息共享能力并进行融合，通过词向量编码器分别得到拼接外部主体类型特征的词向量特征表示$\{h_1, h_2, \ldots, h_n, t | h_i \in \mathbf{R}^{d\times 1}\}$以及词向量中对应关系的关系向量特征表示$\{r_1, r_2, \ldots, r_n | r_i \in \mathbf{R}^{d\times 1}\}$。同时通过利用特征强化编码器中的 RetNet 网络保留机制和自注意力机制的结合来产生较好的上下文特征捕捉性能，同时更好地吸收来自实体类型以及关系类型的特征，从而得到融合关系特征和实体类型的增强级语义信息。具体公式为：

$$S_n = \gamma S_{n-1} + K_n^T V_n \tag{1}$$

$$\text{Retention}(h_n + r_n) = Q_n S_n, \quad n = 1,2,\ldots,x \tag{2}$$

1.2 序列标注模块

本文中序列标注模块（sentence tagging）对两个序列分别标记主体和客体，通过融合关系特征语义编码增强器得出的融合关系特征句子向量以此来预测对象头尾标记的头尾概率，为了更加准确地预测出实体该部分采用 BIO 标注进行标记对应主客体位置，对于该模块具体公式为：

$$P_{i,j}^{\text{sub}} = \text{Softmax}(W_{\text{sub}}(O_n) + b_{\text{sub}}) \tag{3}$$

$$P_{i,j}^{\text{obj}} = \text{Softmax}(W_{\text{obj}}(O_n) + b_{\text{obj}}) \tag{4}$$

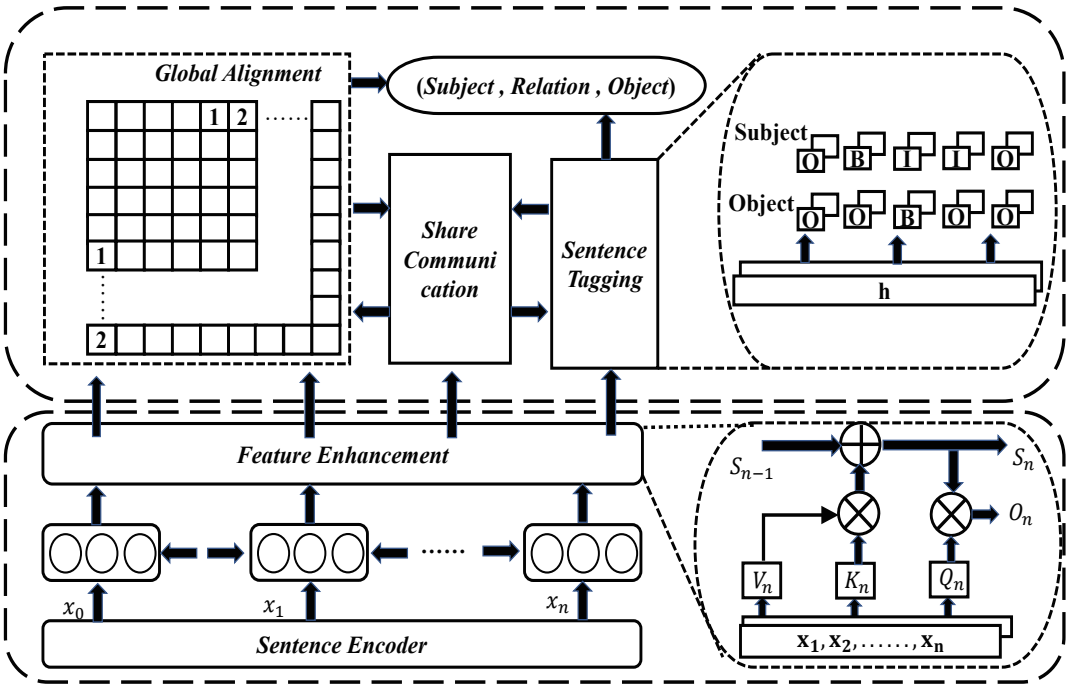

图 1 FRGC 模型结构图

1.3 全局对齐模块

本文中全局对齐模块是对于序列标注以后进行预测，通过使用全局对应矩阵来确定正确的主语和宾语对。首先，本模块会将来自语义特征增强器的主体特征与客体特征进行拼接融合，然后通过该模块会得到对应的全局矩阵得分，对于全局矩阵的分数会有一个特定的阈值，当超过该阈值，对应的矩阵得分就会被保留，反之，矩阵分数将被删除。最后得出一个

二维的得分矩阵来进行标记主客体的一个对齐情况。该模块具体表示为：

$$p_g = \text{Soft max}\,(W_g[O_i^{\text{sub}} + O_i^{obj}] + b_g) \tag{5}$$

式中：P_g 为全局对齐矩阵的概率；W_g 为全局对齐矩阵可训练权重；O_i^{sub}与O_j^{obj}分别为一句话中由增强器得出的对应的主体特征向量和客体特征编码表示向量；b_g 为该模块训练的偏置。

1.4 共享组件模块

近几年主流的联合关系抽取模型普遍存在信息交互不平衡的问题，基于此，本文提出了一种共享交互组件来解决分步走式关系抽取问题存在的问题，通过采用共享组件模块来加强序列标注与全局对齐组件之间的联系，从而解决由于抽取顺序导致的抽取错误问题，首先，通过接收来自特征编码器模块所得到的增强级特征向量，对特征向量进行分为全局对齐区、序列标注区 . 共享交互区 3 个部分，对于共享交互区分别与全局对齐区和序列标注区进行组合映射分别得到对齐特征向量和序列标注信息，以此来促进全局对其区和序列标注区的信息交互与信息共享能力。具体公式为：

$$e = \text{cum max}\,(\text{Linear}([O_t; h_{t-1}])) \tag{6}$$

$$r = 1 - \text{cum max}\,(\text{Linear}([O_t; h_{t-1}])) \tag{7}$$

$$s = \tanh\,(\text{Linear}([O_t; h_{t-1}])) \tag{8}$$

$$P_{s,c_{t-1}} = e_{c_{t-1}} \times r_{c_{t-1}} \tag{9}$$

$$P_{e,c_{t-1}} = e_{c_{t-1}} - P_{s,c_{t-1}} \tag{10}$$

$$P_{r,c_{t-1}} = r_{c_{t-1} - P_{s,c_{t-1}}} \tag{11}$$

$$H_e = P_e + P_s \tag{12}$$

$$H_r = P_r + P_s \tag{13}$$

2 相关实验及分析

2.1 参数设置

本文通过实验，探索了 BERT 模型在自然语言处理任务中的应用。经过反复实验和调查，最终确定了实验设置。具体来说，对于 BERT 模型，隐藏层维度为 768，最大文本长度为 100，批处理大小分别为训练集 64、验证集 24 和测试集 24。训练轮数为 150，丢弃率为 0.3，学习率为 0.000 01，优化器选用 BertAdamW。这些设置经过反复实验和验证，证明了在该任务中 BERT 模型的有效性和稳定性。

2.2 实验及结果

本文采用 3 个关键指标来全面评估实验的性能，分别是精确率（Precision）、召回率（Recall）以及 F_1 值（F_1-score）对模型性能进行评价，公式为：

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$F_1 = \frac{2PR}{P + R} \tag{16}$$

本文采用了 6 个主流文本关系抽取模型进行对比，其中各个算法模型的参数设置与原论文中的设置一致。对比算法为：

PRGC：将关系抽取模型分解为两个部分，先通过判断潜在关系再进行实体抽取以此来解决之前模型的关系冗余问题，提高模型的准确率。

TPLinker[13]：将标注框架统一为字符对链接问题以此来解决重叠关系和暴露误差问题。

SPN[14]：通过将联合抽取视为一个集合预测问题，同时采用分类模型 FFN 来进行实体关系抽取。

CASREL：采用联合抽取的方式进行关系抽取，通过参数共享的层叠指针网络以此来解决三元组 SEO、EPO 问题。

PFN[15]：采用过滤网络，将关系抽取和实体抽取中的有益特征进行强化，对于其中有害的信息进行过滤。

ONEREL：该模型通过一个统一的模块，可以在一步内从非结构化文本中提取所有的三元组，减少因级联误差传播而导致的抽取错误。

实验结果如表 1、表 2 所示。模型（FRGC）在各项指标上对比主流的关系抽取模型均有提升。同时表明，模型（FRGC）在 Duie 数据集上具有较高有效性。

表 1 NYT 数据集实验结果

	P/%	*R*/%	F_1/%
TPLinker	91.4	92.6	92.0
SPN	93.3	91.7	92.5
CASREL	89.7	89.5	89.6
PFN	—	—	92.4
PRGC	93.5	91.9	92.7
ONEREL	93.2	92.6	92.9
FRGC	93.4	92.8	93.1

表 2 Duie 数据集实验结果

	P/%	*R*/%	F_1/%
TPLinker	78.8	71.0	74.7
SPN	77.9	67.5	72.3
CASREL	73.5	79.3	76.3
PFN	—	—	—
PRGC	75.3	76.4	75.8
ONEREL	73.5	76.0	74.7
FRGC	76.9	78.7	77.8

实验结果表明，本文提出的多特征融合文本抽取模型在中英文数据集上取得了优异的性能，其 F_1 指标均超过了现有的先进关系抽取模型。在 Duie 和 NYT 数据集上的对比实验分析中，本文模型在中文 Duie 数据集上比主流模型 CASREL 的 F_1 值提高了 1.5%，同时在 NYT 数据集上也表现出更好的效果。这充分证明了本模型在中英文数据集上的高效性。同时，实验结果也发现 FRGC 模型在英文数据集上的效果提升相对较小。通过分析，认为这是因为 NYT 数据集没有实体类型标注，因此本实验没有在 NYT 数据集上加入实体类型特征，而仅仅加入了关系类型特征，从而导致 FRGC 模型在英文数据集上的提升有限。

3 结语

本文提出了一种创新的融合多特征的模型（FRGC），该模型有效的利用词向量与关系特征作为输入，通过 RetNet 网络进行多特征的交互与融合，有效增强了模型在不同任务间的泛化能力。针对下游适配问题，本文设计了交互式特征过滤机制，通过共享实体对齐与关系解码模块之间的信息，构建动态参数调整策略，以提升模型解码的准确性。实验结果表明，在 Duie 与 NYT 数据集的评估显示，本模型在关系抽取任务上均取得更好的抽取效果，充分验证了本模型在联合关系抽取领域的先进性。

参考文献：

[1] YAN Y, SUN H L, LIU J. A review and outlook for relation extraction[C]//Proceedings of the 5th International Conference on Computer Science and Application Engineering. NewYork:ACM,2021:1-5.

[2] YANG Y, WU Z L, YANG Y X, et al. A survey of information extraction based on deep learning[J]. Applied sciences, 2022, 12(19): 9691.

[3] 鄂海红，张文静，肖思琪，等．深度学习实体关系抽取研究综述 [J]. 软件学报，2019, 30(6): 1793-1818.

[4] NAYAK T, MAJUMDER N, GOYAL P, et al. Deep neural approaches to relation triplets extraction: A comprehensive survey[J]. Cognitive computation, 2021, 13(8): 1215-1232.

[5] ZHAO X, DENG Y, YANG M, et al. a comprehensive survey on relation extraction: recent advances and new frontiers[J]. ACM computing surveys, 2024, 56(11): 1-39.

[6] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[DB/OL].(2022-06-20)[2024-03-19].https://doi.org/10.48550/arXiv.1909.03227.

[7] WANG X, ZHOU W K, ZU C, et al. Instructuie: Multi-task instruction tuning for unified information extraction[DB/OL]. (2023-04-17)[2024-03-19].https://doi.org/10.48550/arXiv.2304.08085.

[8] SHANG Y M, HUANG H Y, MAO X L. Onerel:joint entity and relation extraction with one module in one step[DB/OL].(2022-03-17)[2024-06-12].https://doi.org/10.48550/arXiv.2203.05412.

[9] ROY A, AHMED F, ABDULLAH R, et al. RetNet: retinal disease detection using convolutional neural network[C/OL]// 2023 International Conference on Electrical, Computer and Communication Engineering(ECCE).Piscataway: IEEE, 2023[2024-07-11].https://ieeexplore.ieee.org/document/10101661.

[10] TANG W, XU B F, ZHAO Y Y, et al. UniRel: unified representation and interaction for joint relational triple extraction[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.NewYork: ACM, 2022: 7087-7099.

[11] LI S J, HE W, SHI Y B, et al. Duie: a large-scale chinese dataset for information extraction[C]//8th CCF International Conference. Berlin: Springer, 2019:791-800.

[12] ZHENG H Y, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Brussels: ACL, 2021:6225-6235.

[13] WANG Y C, YU B W, ZHANG Y Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[DB/OL].(2020-10-26)[2024-04-04].https://doi.org/10.48550/arXiv.2010.13415.

[14] SUI D B, ZENG X R, CHEN Y B, et al. Joint entity and relation extraction with set prediction networks[J]. IEEE transactions on neural networks and learning systems, 2023, 35(9): 12784-12795.

[15] YAN Z H, ZHANG C, FU J L, et al. A partition filter network for joint entity and relation extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Brussels:ACL, 2021:185-197.

【作者简介】

解文康 (1999—)，男，江西九江人，硕士，研究方向：自然语言处理。

（收稿日期：2024-10-21）