面向工艺文本的实体与关系最近邻联合抽取模型

杨丹清忻　王裴岩*　徐立军
沈阳航空航天大学计算机学院，辽宁沈阳110134
wangpy@sau.edu.cn
yangdanqingxin@stu.sau.edu.cn
Xlj@ge-soft.com

摘要

该文研究工艺文本中实体关系联合抽取问题，提出了最近邻联合抽取模型（NNJE）。NNJE利用工艺文本中实体边界字间搭配规律建模外显记忆，通过最近邻方法在某种指定关系下为待预测组合检索出具有相似字间搭配的实例，为实体边界识别以及实体对组合提供更有力的限制条件，提升模型预测准确率，改善模型性能。实验设置了工艺文本关系数据集。实验结果表明，该文方法较基线模型准确率P值提高了3.53%，F1值提升了1.03%，优于PURE、CasRel、PRGC与TPlinker等方法，表明提出的方法能够有效地提升三元组抽取效果。

关键词： 工艺文本；实体关系联合抽取；最近邻

Nearest Neighbor Joint Extraction Model for Entity and Relationship in Process Text

Danqingxin Yang　Peiyan Wang　Lijun Xu
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110134, China
wangpy@sau.edu.cn
yangdanqingxin@stu.sau.edu.cn
Xlj@ge-soft.com

Abstract

To address the issue that joint extraction of entity and relationships in process texts, this paper proposes the Nearest Neighbor Integrated Extraction model (NNJE). NNJE uses the rules of word combinations in the process text to model explicit memory and retrieves combinations with similar collocation for the predicted combinations under some specified relationship through nearest-neighbor search. NNJE provides more powerful constraints for entity boundary recognition and entity pair combination, improves the model prediction accuracy and improves the model performance. The experiment set up a joint extraction data set of entity relationships in the process text. Experimental results show that NNJE improves the P score by 3.53% and the F1 score by 1.03% compared with the baseline model,and NNJE is better than PURE, CasRel, PRGC and TPlinker. It shows that the proposed method effectively improves the prediction results of entity-relation triples and improves model performance.

Keywords: process text , joint extraction , nearest neighbor

* 通讯作者

基金项目:辽宁省应用基础研究（2022JH2/101300248）国家自然科学基金（U1908216）

1 引言

实体关系联合抽取是自然语言处理、知识图谱构建的重要任务之一，其目的是从给定的非结构化文本中以（主体，关系，客体）的形式抽取实体对以及它们之间的关系。近年来，知识图谱在制造领域的广泛应用，使得工艺文本实体关系抽取任务尤为重要。

开展工艺文本实体关系联合抽取研究，相较于通用领域更为复杂。工艺文本中包含大量零件、材料特性、设备参数等专业领域知识信息，实体之间关系更为复杂。与通用领域人名、地名与机构名3类实体相比，工艺文本实体类粒度更细（18类），且工艺文本关系类数量远超于通用领域，如Tacrev数据集 (Alt et al., 2020)具有40种关系，SemEval (Hendrickx et al., 2010)数据集具有17种关系，而工艺文本具有62种关系，实体关系联合抽取难度更高。

实体关系联合抽取代表方法有DirectRel (Shang et al., 2022b)、TPLinker (Wang et al., 2020)与OneRel (Shang et al., 2022a)等。上述方法将实体关系联合抽取问题建模为二分图连接判断问题或实体边界字搭配判断问题。DirectRel模型把关系三元组抽取任务看做一个候选实体构成的二分图的连接问题，通过枚举所有小于定长的连续字符作为实体候选集合，该方式会产生过多的负样本，导致准确率较低。特别在工艺文本中实体的平均长度较通用领域（1.85个字符）更长，平均实体长度达3.85个字符（最长实体包含49个字符）。TPLinker与OneRel将三元组抽取问题建模为在特定关系下实体边界字搭配判断问题，这种填表方法改善了基于跨度的方法中实体抽取效果不佳的问题。但TPLinker只将关系类型视为可训练嵌入，实体和关系之间的相互约束不足，导致在组装三元组时效果不佳。OneRel将实体关系联合抽取作为细粒度三元组分类任务，相较于TPLinker采用更轻量级的填表方式。这种填表方式将实体边界进行搭配，但在工艺文本中具有实体边界字搭配的关系兼类歧义现象，导致OneRel抽取的实体关系三元组过度冗余，准确率较低。专业领域中实体关系联合抽取需要以该领域的知识为依据，并兼顾其语言规律特性，使得专业领域的实体关系联合抽取具有一定的难度。工艺文本实体间具有区别于通用领域的特殊字间搭配规律，在数据集规模小的情况下，势必会造成学习的困难。

为解决上述问题，本文提出一种最近邻联合抽取模型，用于工艺文本实体关系联合抽取任务。该方法利用工艺文本中实体边界字间搭配规律建模外显记忆，通过最近邻方法（Nearest Neighbor）为待预测组合检索出具有相似字间搭配的实例，将搭配错误的三元组进行删除以提高模型抽取准确率。实验结果表明，该文方法在工艺文本关系数据集下较基线模型OneRel的准确率P值提高了3.53%，F1值提升了1.03%，优于PURE (Zhong and Chen, 2021)、CasRel (Wei et al., 2020)、PRGC (Zheng et al., 2021)与TPlinker等方法。

KNN是一种基于实例学习的分类算法，适合多分类任务，不需要构建分类模型。有研究指出预训练语言模型在极低的资源条件下表现不稳定，因为在训练期间复杂的示例不容易被记忆到模型的嵌入（Embedding）向量中 (Chen et al., 2022)。目前，在机器翻译 (Zhu et al., 2023)，语言模型 (He et al., 2021)，关系抽取 (Wan et al., 2022)等任务中有加入了KNN外显记忆增强的方法，用来增强模型的鲁棒性与泛化能力。但在联合抽取，尤其是中文领域联合抽取还未见有所研究。

2 相关工作

实体关系联合抽取通常采用以下三种方法：基于参数共享的方法、基于联合解码的方法以及单模块单步骤方法。

2.1 基于参数共享的方法

参数共享的方法将实体关系联合抽取任务分解成实体识别和关系抽取两个子任务，子任务间共享编码层信息以增强交互。

Miwa等 (2016)提出基于端到端的树形结构联合抽取模型，首次采用神经网络的方法实现联合抽取。Zheng等 (2017a)提出了一种混合神经网络模型来提取实体及其语义关系，长短时记忆网络（Long Short Term Memory，LSTM）用于解码实体信息，卷积神经网络用于实现关系抽取。Bekoulis等 (2018)使用CRF层将实体识别任务和关系抽取任务建模为一个多头选择问题，关系抽取层对实体间每种关系做二分类预测。Yu (2020)等人提出了基于分解策略的联合抽取方法，先识别出所有的头部实体，然后对于每个主体识别出相关的客体及关系；最后把这两个任务转化成序列标注问题。Zeng等 (2018)提出了一种端到端神经模型，采用复制机制从句子中提取关系，解码器依次产生关系和主客体。Zeng等 (2020)提出了一个带有复制机

制的多任务学习框架。编码部分使用双向长短时记忆网络建模句子上下文信息，解码部分使用注意力机制与LSTM相结合建模，通过一个全连接层获取输出，结合复制机制生成多对三元组。Zheng等 (2021)为改进信息冗余和扩展性问题，提出一个新的端到端的框架PRGC,将实体关系联合抽取分解成了三个子任务：关系判断、实体抽取和头尾实体对齐，但该模型由于解码步骤复杂存在误差传递和暴露偏差问题。

2.2 基于联合解码的方法

基于联合解码的方法将实体识别和关系抽取两个任务映射在统一的框架结构下，进行全局的优化以及联合解码，加强模块之间的交互，实现实体关系的信息共享。Zheng等 (2017b)提出了一种基于序列标注的联合解码模型，在原有BIES标注方案上进行了扩展，使用BIEOS标签来指代实体的位置，并将关系类型融入标签中。Dai等 (2019)提出一种基于注意力机制的实体关系联合抽取模型，设计了一种可以同时表示实体类型和重叠关系的标记方案。Sui等 (2023)提出了一个非自回归解码的方法，使得各个三元组信息进行双向交互，并融合句子特征向量，用二分图匹配损失函数，针对无序集合的损失计算方式，从各种排列方式中找出和目标匹配损失最小的组合。

2.3 单模块单步骤方法

Shang等 (2022b)通过使用二分图实现对三元组的单步骤抽取。但该模型在数据处理时，生成候选实体集合的方式会产生过多的负样本，对实体平均长度较长的数据会使计算复杂性更高，导致准确率较低。Shang等 (2022a)提出直接从文本语句中抽取三元组，将联合抽取看作细粒度的三元组分类问题。但工艺文本实体和关系粒度较细，实体边界字搭配的关系兼类歧义现象多，导致OneRel抽取的实体关系三元组过度冗余，准确率较低。单模块单步骤模型通过全局矩阵或是图的方法实现从文本中直接抽取三元组，相较于参数共享和联合解码方法避免了级联错误的发生，针对工艺文本关系数据集中实体和关系较多，具有实体边界字搭配的关系兼类歧义现象的问题，单模块单步模型表现仍有不足。

3 最近邻联合抽取模型

3.1 问题定义

给定输入文本$S=\{x_1,x_2,x_3,\ldots,x_L\}$ 以及关系集合$R=\{r_1,r_2,r_3,\ldots,r_N\}$，实体关系联合抽取任务的目标是从给定文本中抽取出所有可能的三元组$T=\{(s_i,r_i,o_i)\}_{i=1}^{K}$，其中$L$为句子长度，$N$为关系数量，$K$为句子中三元组数量，$s$为主体实体，$o$为客体实体。

3.2 模型总框架

本文所提出的最近邻联合抽取模型NNJE（Nearest Neighbor Joint Extraction Model）沿袭了OneRel模型的标注方法：特定关系角标记法（Relation Specific Horns Tagging）。模型总框架如图 1所示，NNJE由OneRel和KNN模块组成，其中预训练语言模型BERT是Bert-base-chinese[0]模型。首先将序列$S=\{x_1,x_2,x_3,\ldots,x_L\}$输入到OneRel模型中，用BERT作为编码器捕捉每个字符的表征向量$\{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3,\ldots,\boldsymbol{e}_L\}$，并枚举出所有可能的$(\boldsymbol{e}_i,\boldsymbol{r}_n,\boldsymbol{e}_j)$组合。由两个全连接层构建出分数向量$\boldsymbol{v}_{(w_i,r_n,w_j)}$，将分数向量$\boldsymbol{v}$输入Softmax函数得到概率$P_{\text{OneRel}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$。同时将分数向量$v$作为查询向量输入到KNN模块中，KNN模块通过欧式距离函数$d(q,h)$计算出与待预测组合具有相似字间搭配规律的前k个键-值对，并计算概率$P_{\text{KNN}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$。将两个模块概率值加权求和得到最终概率$P_{\text{NNJE}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$，解码得到三元组抽取结果。

本章将分别介绍模型各模块的计算过程，包括角标记法、OneRel模型训练过程以及KNN外显记忆的构建与检索过程。

3.3 OneRel

OneRel将联合抽取任务转换成细粒度的三元组分类问题，用一种简洁的标注方法，特定关系角标记法，来标注三元组信息。这种角标记法从“BIE”(Begin,Inside,End)标注格式演化

[0]https://huggingface.co/bert-base-chinese

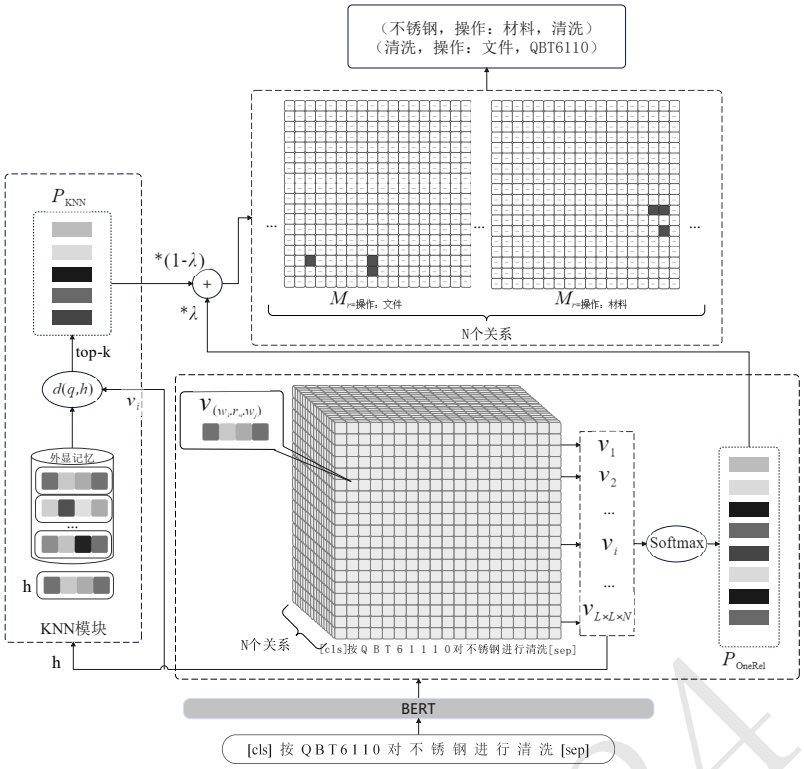


Figure 1: NNJE模型架构图

而来，并用“H”、“E”表示实体的开始以及结束。例如，HB指主体的开始位置，TE指客体的结束位置。OneRel把指定关系下的字间组合定义为四类，用于检测实体边界。(1)HB-TB：指在某种特定关系条件下主体和客体的起始位置标记。(2)HB-TE：指在某种特定关系条件下主体的起始位置和客体的结束位置标记。(3)HE-TE：指在某种特定关系条件下主体和客体的结束位置标记。(3)“-”：指在任何关系下都不构成实体边界的字符对，即除了上述情况下的所有字符对。如图 2所示，在句子“铆钉的铆接验收标准按QBT1111。”中有两个三元组，分别是：（铆钉，操作：文件，QBT1111）和（铆钉，操作：零件，铆接）。其中，在关系“操作：文件”下，主体和客体分别是“铆钉”和“QBT1111”。那么组合（铆，操作：文件，Q）就在子矩阵$\boldsymbol{M}_{r=\text{操作：文件}}$中被标记为“HB-TB”。同样的，组合（铆，操作：文件，1）被标记为“HB-TE”，组合（钉，操作：文件，1）被标记为“HE-TE”。不同关系将被标记在不同的子矩阵中。这种标记方法使得一个序列中所有的$(\boldsymbol{e}_i, \boldsymbol{r}_n, \boldsymbol{e}_j)$组合都能用一个三维张量$\boldsymbol{M}^{L\times N\times L}$表示。由于这种方法能够同时表示出主客体边界以及主客体之间的关系，因此解码时可以直接从$\boldsymbol{M}^{L\times N\times L}$中解码得到三元组。

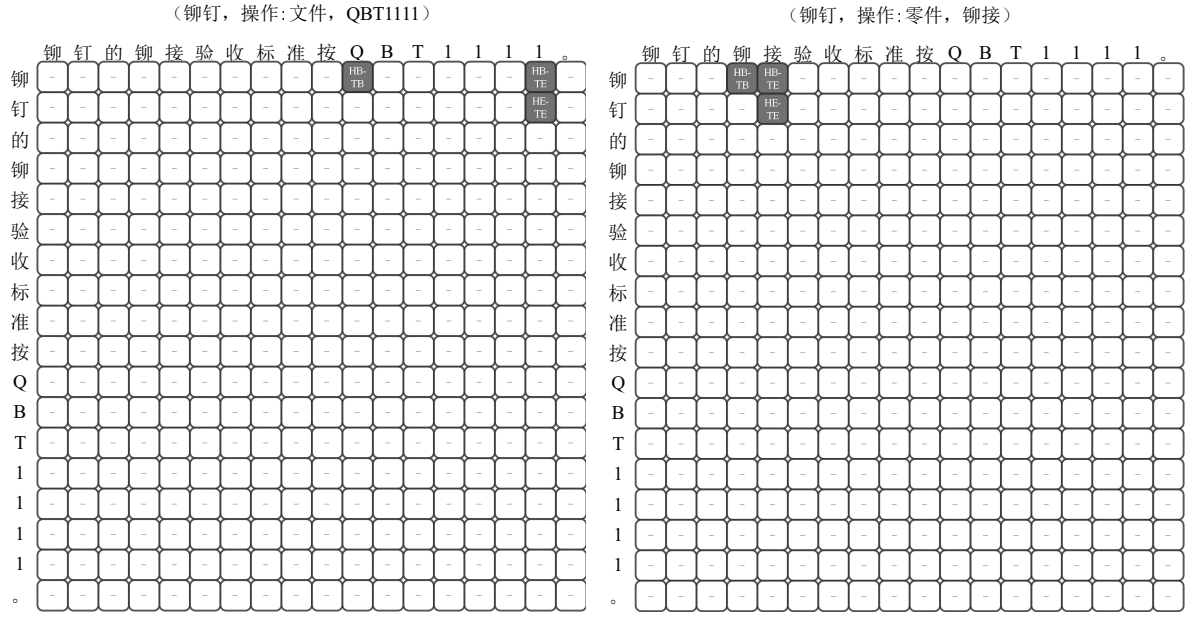
OneRel模型首先将序列输入预训练语言模型BERT中进行编码，如式(1)。

$$\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_L\} = BERT\left(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L\}\right) \tag{1}$$

然后枚举出所有可能的$(\boldsymbol{e}_i, \boldsymbol{r}_n, \boldsymbol{e}_j)$组合,通过一个分类器为每个组合分配高置信度标签。受知识图谱嵌入技术 (Nickel et al., 2016)的启发，该分类器使用两个全连接网络实现并行评分，可以自适应地学习从实体特征到实体对表示的映射函数，将分数函数定义为如式(2)。

$$\boldsymbol{v}_{(w_i,r_n,w_j)_{n=1}^{N}} = \boldsymbol{R}^T \phi(\text{drop}(\boldsymbol{W}[\boldsymbol{e}_i; \boldsymbol{e}_j]^T + \boldsymbol{b})) \tag{2}$$

其中，[;]表示拼接操作，$\boldsymbol{W} \in \mathbb{R}^{d_e \times 2d}$，$d_e$表示实体对表征的维度，$b$是可训练的权重。drop(·)是Dropout函数 (Srivastava et al., 2014)用来防止过拟合，$\phi(\cdot)$为RELU激活函数。$\boldsymbol{R} \in \mathbb{R}^{d_e \times 4N}$，其中4是角标记标签的数量，$\boldsymbol{R}$和$\boldsymbol{W}$均可以视为可训练的权重。$\boldsymbol{v}$是模型输出的分数向量。



(a) $M_{r=\text{操作：文件}}$角标记示例　　(b) $M_{r=\text{操作：零件}}$角标记示例

Figure 2: 角标记关系示例图

在OneRel模型预测时，将得到的分数向量输入到Softmax函数中，如式(3)，得到(w_i, r_n, w_j)组合的概率分布$P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S})$。

$$P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) = \text{Softmax}(\boldsymbol{v}_{(w_i,r_n,w_j)}) \tag{3}$$

损失函数定义如式(4)。

$$\mathcal{L}_{\text{triple}} = -\frac{1}{L \times N \times L} \times \sum_{i=1}^{L}\sum_{n=1}^{N}\sum_{j=1}^{L} \log P(y_{(w_i,r_n,w_j)} = t_{(w_i,r_n,w_j)}|\boldsymbol{S}) \tag{4}$$

其中，$t_{(w_i,r_n,w_j)}$为从数据集中获取的正确标签。

3.4 KNN

KNN的外显记忆数据是以键-值对的形式存在的，在KNN模块中，将训练数据中的序列输入到OneRel模块中，通过模型训练得出的每个(w_i, r_n, w_j)组合的分数向量$v_{(w_i,r_n,w_j)}$作为外显记忆集D中的键h。值u则是由每个(w_i, r_n, w_j)组合对应的角标记标签构成，如式(5)。

$$D = \{(\boldsymbol{h}, u)|(\boldsymbol{v}_{(w_i,r_n,w_j)}, t_{(w_i,r_n,w_j)})\} \tag{5}$$

KNN模块并不参与训练，仅利用实例之间的相似性进行预测完成分类任务。在模型预测阶段，将模型输出的分数向量$\boldsymbol{v}_{(w_i,r_n,w_j)}$作为KNN模块的输入，记作查询向量$v_i$，计算$v_i$与外显记忆集中每个键$h$的欧氏距离，取出与查询向量$\boldsymbol{v}_i$距离最小的$k$个键-值对，构成查询向量的最近邻集合$D^i$，以投票机制计算出当前查询向量的标签概率分布，如式(6)。

$$P_{\text{KNN}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) \propto \sum_{(\boldsymbol{h},u)\in D^i} \text{I}_{y_i=u} \exp(\frac{-d(\boldsymbol{v}_i, \boldsymbol{h})}{T}), T \in R^+ \tag{6}$$

模型最终加权求和OneRel和KNN模型预测概率分布，如式(7)。

$$P_{\text{NNJE}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) = (1-\lambda)P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) + \lambda P_{\text{KNN}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) \tag{7}$$

4 实验数据与实验结果分析

4.1 实验数据

本文使用了人工标注的工艺文本关系数据集，数据来源为某型号飞机所遵照的工艺规范文件。实验数据集包含4421条句子，人工制定了62种关系类，共标注7646个关系，按照8:1:1的比例划分数据集为训练集、测试集和验证集。具体关系类及其数量分布情况见表 1。

关系子类	关系数量	关系子类	关系数量	关系子类	关系数量
操作:工艺辅料	830	属性：表注	165	工艺辅料：部位	21
操作：材料	111	属性：文件	101	材料：部位	18
操作：多余物	55	属性：图注	73	孔：部位	22
操作：零件	714	孔：表注	2	工具：部位	28
操作：部位	367	孔：文件	8	零件：部位	152
操作：孔	110	工具：图注	31	部位：属性值	34
零件：零件编号	156	工具：表注	25	零件：属性值	43
材料：材料牌号	23	工具：文件	8	工具：属性值	31
工具：工具编号	8	部位：文件	4	工艺辅料：属性值	10
工艺辅料：工艺辅料编号	771	工艺辅料：表注	22	操作：属性值	178
零件：属性	395	工艺辅料：文件	11	孔：属性值	9
操作：属性	454	零件：表注	6	多余物：属性值	6
工具：属性	31	零件：图注	17	材料：属性值	9
部位：属性	193	零件：文件	6	属性：属性值	796
工艺辅料：属性	169	部位：图注	6	数量：部位	11
孔：属性	94	操作：图注	75	数量：操作	16
材料：属性	46	操作：文件	376	数量：孔	16
部位：多余物	42	操作：表注	24	数量：零件	53
部位：工艺辅料	18	操作-用具	340	数量：工艺辅料	59
部位：零件	6	零件：材料	60	数量：工具	5
文件：表注	4	工具：材料	31		

Table 1: 工艺文本数据集关系类型统计信息

4.2 对比方法

本文选择PURE(Zhong and Chen, 2021)、SpERT(Eberts and Ulges, 2020)、CasRel (Wei et al., 2020)、PRGC (Zheng et al., 2021)、TPLinker (Wang et al., 2020)和OneRel (Shang et al., 2022a)作为对比方法。其中，PURE将任务分成实体识别和关系判断两个子任务模型分别进行训练。SpERT、CasRel与PRGC是基于参数共享的实体关系联合抽取方法。SpERT先抽取出文本中的实体根据抽取出的实体对实体对做关系分类。CasRel采用层叠指针网络的思想，先识别所有可能的主体集合，在每个给定的关系类别下抽取与主体存在该关系的客体。PRGC采用端到端框架，先判断预测句子中可能存在的关系集合，再针对每个关系分别进行两次序列标注操作，分别提取出主体客体，再使用全局关联矩阵确定实体对。TPLinker是基于联合解码的方法，定义了3种链接方式判断实体边界，根据实体边界组合创建关系表格以抽取三元组。OneRel将实体关系联合抽取转换成细粒度的三元组分类问题，用特定关系角标注法标注实体边界，进行三元组抽取。

4.3 评价指标

本文使用准确率P（Precision）、召回率R（Recall）、F1值作为模型的评价指标，F1值越

高代表抽取效果越好。评估方法由下式列出。

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{10}$$

其中，TP表示将正类抽取为正类的数量，指的是模型在实体关系联合抽取任务中，对输入实例的主体实体、客体实体以及实体对间的关系类型全部抽取正确的实例数量。相应地，FP为将负类抽取为正类数，FN为将正类抽取为负类数。

4.4 参数设置

本文实验采用了Torch 1.8.0框架。模型的超参数取值如表 2所示。在超参数中，k代表最近邻样本的数量，KNN模块的概率权重用 λ 表示，λ 的取值范围为[0.05,0.95]，本文选取该范围中在验证集上效果最好的 λ 作为最终模型参数取值。

超参数	取值范围
训练轮次（Epoch）	100
训练批大小（Batch Size）	2
学习率（Learning Rate）	5e-6
k	5
λ	[0.05,0.95]

Table 2: 模型超参数设置

4.5 实验结果分析

本文选用当前效果较为突出的6个深度学习模型作为对比实验，在工艺文本关系数据集下进行评估，NNJE模型的F1值优于其他对比方法，结果如表 3所示。通过分析实验结果可进一步得出如下结论。

模型	P	R	F1
TPLinker	37.53%	24.42%	29.58%
SpERT	37.72%	32.77%	35.07%
PURE	49.08%	33.52%	39.83%
CasRel	50.39%	34.18%	40.73%
PRGC	57.30%	**52.30%**	54.70%
OneRel	56.75%	52.20%	54.38%
NNJE	**60.28%**	51.27%	**55.41%**

Table 3: 实体关系联合抽取的实验效果

第二十三届中国计算语言学大会论文集，第406页-第417页，太原，中国，2024年7月25日至28日。
卷1：主会论文

模型	子任务	P	R	F1
OneRel	实体识别	80.21%	71.60%	75.66%
	实体对组成	58.57%	53.76%	56.06%
	关系分类	57.62%	53.00%	55.22%
	三元组	56.75%	52.20%	54.38%
NNJE	实体识别	**81.78%**	69.90%	75.37%
	实体对组成	**62.11%**	52.51%	**56.91%**
	关系分类	**61.07%**	51.94%	**56.13%**
	三元组	**60.28%**	51.27%	**55.41%**

Table 4: 实体关系联合抽取子任务实验效果

首先，NNJE模型F1值比次优模型OneRel高出1.03%，准确率P值提高了3.53%，验证了引入KNN方法能够有效提升单模块单步骤方法的模型性能。

其次，NNJE和OneRel模型抽取结果明显优于TPLinker、PURE和CasRel三个模型，说明在工艺文本实体关系联合抽取任务中，使用全局矩阵的方式优于其他各种分步骤分模块的抽取方式，这是因为全局矩阵能够充分捕捉工艺文本中实体与实体之间的特征关联，更适应工艺文本的复杂性以及专业性。与这两个模型相比，由于工艺文本关系数据集相较通用领域数据集粒度更细，而TPLinker的标注复杂度高，若关系数目大，解码部分矩阵参数量大，从而导致训练收敛速度慢和抽取性能不高等问题；SpERT在未标注数据中随机生成负样本，对数据质量的要求较高，因此在本文数据样本少且有大量混淆的情况下抽取效果差；PURE是两个模型独立训练，受级联错误影响，抽取效果不佳；CasRel在抽取客体时，需计算每个关系下对应的客体，容易导致关系冗余等问题。而本文模型NNJE优于OneRel与其他几种对比方法，验证了NNJE模型引入KNN方法的有效性。

此外，本文在不同子任务上进一步探索了NNJE模型效果，即实体识别、实体对组成以及关系分类任务。如表 4所示，在实体对组成和关系抽取子任务上，NNJE展现了优秀的性能。并且在实体识别F1值相当的情况下，NNJE模型的实体对组成、关系判断子任务的准确率和F1值都显著高于OneRel模型。这证明KNN通过查询与待预测组合有相似字间搭配规律的键-值对，增强了实体对识别以及关系分类性能。

4.6 KNN权值影响分析

本文探索了KNN权值λ对模型效果的影响，λ在[0.05,0.95]间取值，模型效果如图 3所示。从图中可以看到F1值随着λ的增长总体呈上升趋势，达到峰值后有所回落，除极端权值[0.05,0.10]和[0.85,0.95]外，当KNN权值越大时普遍效果更好。

4.7 实例分析

表 5中展示了文本、正确三元组、OneRel方法输出的三元组与NNJE方法输出的三元组，阐述了KNN模块对实体错误和关系分类错误两种错误三元组的删除情况。在表 5的实例1中，操作“压印”与部位“铣切边”和“凸缘边”之间不存在关系，部位“凸缘边”与属性“间距”应存在“部位：属性”的关系，。同样的，在实例2中OneRel模型错误地将铆钉头的零件编号“A-286”识别成工具“铣刀”的工具编号，把“硬质合金刀齿”错误识别成零件，并将不与任何实体有关系的数量实体“25个”和它们组成了三元组，KNN的加入删除了这些错误三元组。

在表 6中还展示了KNN对每个组合在正确标签上的预测概率及KNN检索出的实例等信息，从而进一步对KNN能够删除错误三元组的原因进行了分析。实例中，OneRel方法将零件实体“铆钉”和部位实体“周围的表面”识别成了“零件：部位”的关系。在工艺文本中“零件：部位”关系通常指零件操作时强调在某个部位操作，或零件以及零件本身的部位。在制造与装配工艺中，“钉”这类标准件不会对其“面”有特别操作与关注。在KNN的实例中，“钉”与“面”不会构成“零件：部位”关系下三元组中的实体边界，即主体尾和客体尾，故而将错误三元组删除。

从上述实例中可以看出，KNN能够根据实体边界的构成规律，构建实体边界字间的关系实例，通过检索拥有相似字间搭配规律的实例，有效地删除错误三元组，从而提高抽取准确率。

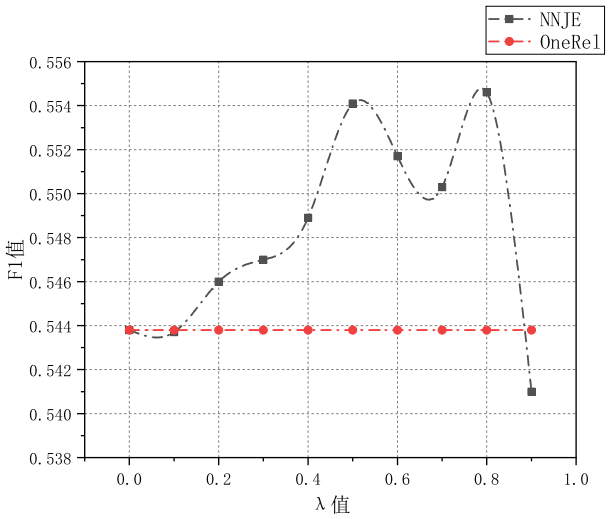


Figure 3: KNN权重变化下的F1值

实例1	
文本	在压印燃油传输槽口时，压印燃油传输槽口至零件边缘之间和压印燃油传输槽口至连接孔之间的最小间距如表6-3和图6-3所示，表6-3中的“E”尺寸适用于槽口至铣切边或凸缘边的间距。
正确三元组	（压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3）（压印，操作：部位，燃油传输槽口至连接孔之间）（“E”尺寸，属性:表注，表6-3）（最小间距，属性:图注，图6-3）（压印，操作：属性，最小间距）（压印，操作：部位，燃油，燃油传输槽口至零件边缘之间）（凸缘边，部位：属性，间距）
OneRel	（压印，操作：部位，凸缘边）（铣切边，部位：属性，最小间距）（凸缘边，部位：属性，最小间距）（最小间距，属性:图注，图6-3）**（压印，操作：部位，铣切边）（压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3）**
NNJE	**（最小间距，属性:图注，表6-3）（最小间距，属性:图注，图6-3）（压印，操作：部位，燃油传输槽口）**（铣切边，属性:图注，最小间距）
删除错误三元组	（压印，操作：部位，凸缘边）（铣切边，部位：属性，最小间距）（凸缘边，部位：属性，最小间距）（最小间距，属性:图注，图6-3）（压印，操作：部位，铣切边）（压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3）
实例2	
文本	铣切A-286铆钉头时应采用ZT507B双向旋转的约有25个硬质合金刀齿的铣刀。
正确三元组	（铣切，操作:零件，铆钉头）（铣刀，工具:工具编号，ZT507B）（铆钉头，零件:零件编号，A-286）
OneRel	（25个，数量:零件，硬质合金刀齿）**（铣切，操作:零件，硬质合金刀齿）**（25个，数量:零件，铆钉头）**（铣切，操作:零件，铆钉头）**（铣刀，工具:工具编号，A-286）（铆钉头，零件:零件编号，A-286）
NNJE	**（铣切，操作:零件，铆钉头）（铆钉头，零件:零件编号，A-286）**
删除错误三元组	（25个，数量:零件，硬质合金刀齿）（铣切，操作:零件，硬质合金刀齿）（25个，数量:零件，铆钉头）（铣刀，工具:工具编号，A-286）

Table 5: 删除错误三元组实例分析

文本	按NACA铆接方法正确安装的铆钉见图6-3，注意铆钉周围的表面要光滑。				
正确三元组	（铆钉，零件:图注，图6-3）				
OneRel	（安装，操作:零件，铆钉）**（铆钉，零件:图注，图6-3）** （铆钉，零件:部位，周围的表面）				
NNJE	（安装，操作:零件，铆钉）**（铆钉，零件:图注，图6-3）**				
（字符，关系，字符，标签）组合	P_{OneRel}	P_{NNJE}	检索出组合及标签		检索出组合所在句子
			组合	标签	
（“钉”，“零件:部位”，“面”，“-”）	0.402	0.56	（“纸”，“工艺辅料：工艺辅料编号”，“L”）	-	如果表面用QBT6520揩布 或QBT640中性牛皮**纸**进行了保护，则允许在24小时内完成硅底胶或CMS-S**L**系列胶粘剂的施工。
			（“条”，“工艺辅料：工艺辅料编号”，“7”）	HE-TE	用真空袋密封胶**条**（QBT5537、QBT5537或QBT5537）将真空袋（QBT5664、QBT5664或QBT5664）密封到模具上，确保真空袋不架桥。
			（“胶”，“工艺辅料：工艺辅料编号”，“1”）	-	如有凸起或阶差允许 用QBT923（150目或更细）砂纸打磨修平，拼接缝内若有凹坑，允许使用CMS-AD-106发泡**胶**进行修补，修补后拼接缝及泡沫表面应用可剥布（QBT6000，QBT600**1**）或脱模布（QBT3463）铺贴覆盖。
			（“件”，“工艺辅料：工艺辅料编号”，“理”）	-	应力消除：通常，工程图样规定焊后进行热处**理**的低合金钢和碳钢零件的闪光焊缝以及不锈钢零**件**的闪光焊缝不要求消除应力。但是HY-Tuf钢（高韧钢），4140，4340和300M低合金钢零件则要求在焊后冷却至93℃～204℃(200～400)温度范围后进行消除应力，其消除应力的方法，可以用闪光焊机来加热焊缝或用感应线圈将其加热到538℃～649℃(1000～1200)的温度范围，也可以按QBT5000消除应力。

Table 6: KNN检索实例分析

5 总结

本文利用工艺文本中实体边界字间搭配规律，提出面向工艺文本实体与关系的最近邻联合抽取模型NNJE。实验结果表明：首先，本文提出的最近邻联合抽取模型NNJE效果优于对比方法，表明了NNJE模型的优越性；其次，从NNJE模型与单模块单步骤模型OneRel的对比结果看，KNN能够有效增强抽取效果，且当KNN权值越高时，效果越好；再次，分别在各子任务上研究了NNJE模型，在实体对识别以及关系分类子任务上均有着较好的抽取效果；最后，通过实例分析，NNIE模型能够根据实体边界的构成规律，构建实体边界字间的关系实例，为待预测组合检索出与之字间搭配相似的键-值对。未来工作中，将对填表方式进行压缩，减少由于关系类数量过多导致产生过多负样本，构建更适用于多关系类别数据的抽取模型。

参考文献

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1558–1569. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation extraction as open-book examination: Retrieval-enhanced prompt tuning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2448.

Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. 2019. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6300–6308.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5703–5714. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 33–38. The Association for Computer Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022a. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Yuming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. 2022b. Relational triple extraction: One step is enough. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4360–4366. ijcai.org.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Jiwei Li. 2022. Rescue implicit and long-tail cases: Nearest neighbor relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1731–1738. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1476–1488. Association for Computational Linguistics.

Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Tingwen Liu, Yubin Wang, Bin Wang, and Sujian Li. 2020. Joint extraction of entities and relations based on a novel decomposition strategy. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2282–2289. IOS Press.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017a. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017b. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1227–1236. Association for Computational Linguistics.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. PRGC: potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6225–6235. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.

Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023. INK: injecting knn knowledge in nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15948–15959. Association for Computational Linguistics.