软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007219]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向低资源关系抽取的自训练方法*

郁俊杰[1], 王 星[2], 陈文亮[1], 张 民[1]

[1](苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

[2](腾讯 AI Lab, 广东 深圳 518000)

通信作者: 陈文亮, E-mail: wlchen@suda.edu.cn

摘 要: 自训练是缓解标注数据不足问题的常见方法, 其通常做法是利用教师模型去获取高置信度的自动标注数据作为可靠数据. 然而在低资源场景关系抽取任务上, 该方法不仅存在教师模型泛化能力差的问题, 而且受到关系抽取任务中易混淆关系类别的影响, 导致难以从自动标注数据中有效地识别出可靠数据, 同时产生大量难以利用的低置信度噪音数据. 因此, 提出一种有效利用低置信度数据的自训练方法 ST-LRE (self-training approach for low-resource relation extraction). 该方法一方面基于复述增强的预测方法来加强教师模型筛选可靠数据的能力; 另一方面, 基于部分标注模式从低置信度数据中提炼出可利用的模糊数据. 基于模糊数据的候选类别集合, 提出了基于负标签集合的负向训练方法. 最后, 为了支持可靠数据和模糊数据的融合训练, 提出一种支持正负向训练的联合方法. 在两个广泛使用的关系抽取数据集 SemEval2010 Task-8 和 Re-TACRED 的低资源场景上进行实验, ST-LRE 方法取得显著且一致的提升.

关键词: 自然语言处理; 信息抽取; 关系抽取; 低资源; 自训练

中图法分类号: TP18

中文引用格式: 郁俊杰, 王星, 陈文亮, 张民. 面向低资源关系抽取的自训练方法. 软件学报. http://www.jos.org.cn/1000-9825/7219.htm

英文引用格式: Yu JJ, Wang X, Chen WL, Zhang M. Self-training Approach for Low-resource Relation Extraction. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7219.htm

Self-training Approach for Low-resource Relation Extraction

YU Jun-Jie[1], WANG Xing[2], CHEN Wen-Liang[1], ZHANG Min[1]

[1](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

[2](Tencent AI Lab, Shenzhen 518000, China)

Abstract: Self-training, a common strategy for tackling the annotated-data scarcity, typically involves acquiring auto-annotated data with high confidence generated by a teacher model as reliable data. However, in low-resource scenarios for Relation Extraction (RE) tasks, this approach is hindered by the limited generalization capacity of the teacher model and the confusable relational categories in tasks. Consequently, efficiently identifying reliable data from automatically labeled data becomes challenging, and a large amount of low-confidence noise data will be generalized. Therefore, this study proposes a self-training approach for low-resource relation extraction (ST-LRE). This approach aids the teacher model in selecting reliable data based on prediction ways of paraphrases, and extracts ambiguous data with reliability from low-confidence data based on partially-labeled modes. Considering the candidate categories of ambiguous data, this study proposes a negative training approach based on the set of negative labels. Finally, a unified approach capable of both positive and negative training is proposed for the integrated training of reliable data and ambiguous data. In the experiments, ST-LRE consistently demonstrates significant improvements in low-resource scenarios of two widely used RE datasets SemEval2010 Task-8 and Re-TACRED.

Key words: natural language processing; information extraction; relation extraction; low-resource; self-training

---

* 基金项目: 国家自然科学基金 (62376177, 61936010)
收稿时间: 2023-10-10; 修改时间: 2024-01-18; 采用时间: 2024-04-19; jos 在线出版时间: 2024-07-03

关系抽取 (relation extraction) 是自然语言处理的一项基础任务, 其旨在获取给定句子中两个实体之间的语义关系[1]. 近年来, 随着“预训练-微调”范式的快速发展[2], 将预训练模型以微调形式应用到关系抽取任务上的方法取得了显著进展[3]. 然而, 关系抽取任务在实际应用中仍然受到标注数据稀缺问题的困扰. 主要原因是对于大多数基于关系抽取的应用来说, 关系定义与具体应用场景紧密相关[4,5], 很难有大量定制化的标注数据. 同时, 人工标注关系抽取数据是一项耗时耗力且昂贵的工程. 因此, 作为一种替代方案, 自动标注关系抽取数据在研究界和产业界都引起了广泛关注[6–8].

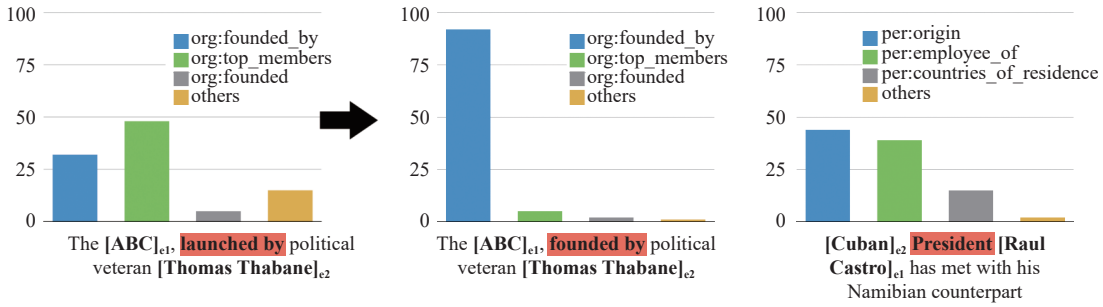
自训练 (self-training) 是一种简单且有效的自动标注方法[9]. 其基本思想是利用少量人工标注数据作为种子数据训练一个教师模型, 然后使用教师模型去自动标注大量数据, 并从中挑选预测置信度高的数据并以硬标注 (hard label) 模式将其作为可靠数据, 即选取预测概率最高的类别作为唯一标签. 最后, 将这些可靠数据和种子数据一起用于训练学生模型. 由于仅需要少量种子数据, 自训练适用于低资源场景下的关系抽取任务, 更接近于真实的应用场景. 在本文中, 我们同样采用自训练框架来提高关系抽取系统在低资源场景下的性能.

在前人的研究工作中, 研究人员通常会选择具有高置信度的自动标注实例作为可靠数据, 同时弃用剩余数据, 这种简单策略在实际应用中取得了一定的成功[10–12]. 在高置信度数据中, 教师模型赋予某个关系远超其他关系类别的预测概率, 并以此关系类别作为唯一的标签 (硬标注模式) 构造可靠数据. 在早期实验中, 我们尝试了这种简单的自训练方案来解决低资源场景下关系抽取任务的语料稀缺问题. 从初步实验结果来看, 该方案存在两个问题限制了自训练性能. (1) 可靠数据筛选受到制约. 由于教师模型是在少量人工标注数据上训练得到的, 其泛化能力较差, 导致一些训练集中未出现的语义表达无法被教师模型识别, 从而影响可靠数据的筛选. (2) 无法充分利用自动标注数据. 除了高置信度的数据可作为可靠数据外, 剩余的大量低置信度数据被弃用.

为了缓解可靠数据筛选受到制约的问题, 本文引入复述技术, 提出复述增强的预测方法帮助教师模型选择部分低置信度数据作为可靠数据. 在本文中, 复述是指在不改变原句语义表达的前提下, 使用新的文本表达组成新句子. 使用复述数据有助于增加未标注句子的表达多样性, 从而提高教师模型的泛化能力. 图 1(a) 中展示了复述增强方法帮助教师模型进行预测的例子 ($[\ ]_{e1}$ 和 $[\ ]_{e2}$ 用于标注头尾实体). 由于原句 (左侧) 中的表达“launched by”未在种子数据中出现过, 教师模型并不能正确地预测出输入句子的关系类别“org:founded_by”. 而原句的一个复述句 (右侧) 使用了更为常见的“founded by”来描述某组织机构被某人创立的语义信息, 因此, 教师模型非常有把握地将其预测为“org:founded_by”的关系. 本文利用大型预训练语言模型 (大语言模型/LLM) 来生成复述数据. 为了提高复述质量, 本文精心设计了面向关系抽取任务的复述生成提示语和基于实体约束的后处理方法. 在复述增强的预测方法中, 我们将原未标注句子 (称为主句) 和生成的复述句子 (称为辅句) 组成句子包. 教师模型对句子包中的所有句子进行标签预测. 当主句不满足高置信度条件时, 检查辅句中是否存在满足条件的预测, 若存在, 则将对应概率分布赋予主句, 并将其以硬标注模式构建为可靠数据.

对于无法充分利用低置信度数据的问题, 通过进一步实验分析, 我们发现一些句子存在语义表达与多种关系类别均有关联的现象. 这类现象使得教师模型感到困惑, 在部分关系类别上给出相近的预测概率, 导致数据无法被使用. 图 1(b) 展示了一个令教师模型感到困惑的低置信度示例: 对于“Cuban President Raul Castro (古巴总统劳尔·卡斯特罗)”的表述, 尽管表述简单但其蕴含着多种可能的语义关系. 从字面意思上来看, 就职单位信息 (per:employee_of) 是最直接的关系. 但是, 由于单位是国家, 职位是总统, 因此一般来说也可以推断出人物原籍信息 (per:origin), 甚至居住国家信息 (per:countries_of_residence). 由于这些模糊关系的存在, 导致教师模型很难区分当前句子具体属于哪一种关系. 在类似情况下, 教师模型可能会对某些容易混淆的关系给出相近的高概率而对其余关系类别给予低概率, 或者教师模型对所有关系都没有把握, 从而对所有关系都打上相近的低概率. 在前人的研究中, 由于低置信度的实例会带来大量的噪音, 进而导致系统的性能下降. 因此, 这类数据通常会被舍弃. 然而, 我们认为直接忽略所有低置信度数据可能并不合适, 因为它们可能包含有用信息且能够帮助到学生模型的训练. 例如, 对于图 1(b) 中的实例, 尽管教师模型并不确定具体应该是哪种关系, 但比较肯定: 答案大概率是前 3 种关系中的一种. 理想情况下, 我们希望能够充分利用所有自动标注的实例来改进关系抽取系统, 但噪音过多的数据的使用

是非常困难的. 因此, 本文将剩余的低置信度数据分成两个集合: 模糊数据和噪音数据. 模糊数据指的是教师模型在某些关系上预测出相近高概率的实例, 而噪音数据则是那些教师模型对所有关系都分配了低概率的实例. 对于模糊数据, 本文做出一个假设: 教师模型尽管不知道哪个关系是确切的答案, 但它确实知道: (1) 答案 (很大概率上) 在候选类别集合中 (类似于图 1(b) 中的前 3 个关系); (2) 答案不在那些概率非常低 (负类别) 的类别集合之中 (类似于图 1(b) 中的“others”). 因此, 不同于硬标注模式来处理高置信度的数据, 本文提出以部分标注 (partial label) 模式处理模糊数据, 即为每个模糊数据提供一个候选类别集合, 称为正标签集合, 而剩余类别则归类到一个负标签集合中. 基于本文假设, 负标签集合中的类别作为答案标签的置信度低, 而相反地作为负标签则有较高的置信度. 因此, 本文利用基于负标签集合的负向训练方法来使用模糊数据训练模型, 从而充分利用低置信度数据.



(a) 复述增强的例子 (b) 模糊关系的例子

图 1 自动标注句子样例

基于上述思路, 本文提出一种基于复述数据和模糊数据增强的自训练方法 ST-LRE (self-training approach for low-resource relation extraction) 来充分利用自动标注数据, 具体的数据使用情况如图 2 所示. 该方法使用常见的自训练方案, 基于少量的种子数据训练教师模型用于数据的自动标注. 但为了提高教师模型的泛化能力, ST-LRE 通过大语言模型的复述能力来扩充未标注数据的语义表达多样性. 同时, 为了充分利用低置信度数据, ST-LRE 基于模糊数据的概念挖掘了低置信度数据中的确定信息, 并提出面向模糊数据的部分标注模式和负向训练方法. 最终, ST-LRE 的联合训练方法融合了可靠数据的正向训练和模糊数据的负向训练. 为了验证 ST-LRE 的能力, 本文在两个广泛使用的关系抽取任务数据集上进行了低资源场景的实验. 实验结果表明, 本文所提方法能够有效解决低资源场景下教师模型泛化差和低置信度数据利用率低的问题, 并最终提升关系抽取系统的性能.

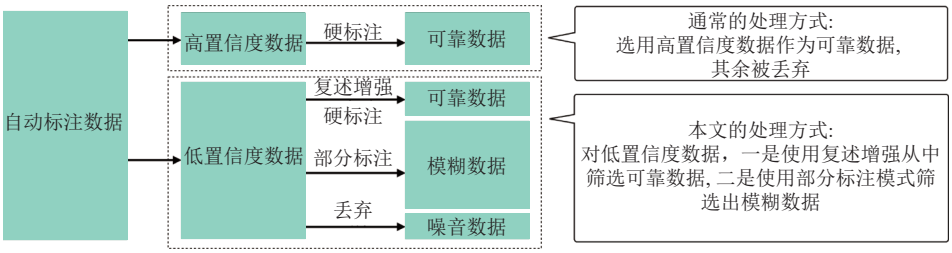


图 2 自动标注数据的处理

本文贡献可归纳为:

(1) 提出了基于复述增强的预测方法. 通过大语言模型的复述能力增加未标注数据的表达多样性, 以此提高教师模型的泛化能力, 进而提高可靠数据的筛选能力.

(2) 提出了一种提取低置信度数据中可用信息的方法. 基于部分标注模式获取候选类别集合的方式建立模糊数据集, 并提出了面向模糊数据的负向训练方法.

(3) 通过两个常用关系抽取数据集上的实验验证了所提自训练方法 ST-LRE 的有效性.

1 相关工作

1.1 低资源关系抽取

近年来, 由于标注数据获取困难的问题, 低资源场景关系抽取任务受到了广泛的关注[13]. 低资源关系抽取通常指的是在仅有少量标注样本的前提下构建关系抽取系统. 由于人工标注费时费力, 如何低成本地扩充标注数据或增强已有标注数据的表达能力成为低资源关系抽取的重点. 为此, 一个方向是引入外部知识来增强已有数据的表达. 比如, 欧阳丹彤等人[14]基于本体的远程监督方式进行样本扩充. 此外, Yang 等人[15]提出引入实体概念知识库中的信息来增强实体的表达能力. 相应地, 引入关系类别的语义信息也是缓解数据匮乏问题的有效方法[16,17]. 另一个方向是基于自动标注的方法来扩大数据样本的规模. 比如, 基于现有的关系知识库, Mintz 等人[6]提出远程监督技术来生成大规模自动标注数据. 相似地, 朱苏阳等人[18]基于中文维基百科中的半结构化知识, 构建了面向家庭关系的数据集. 基于机器翻译系统, 胡亚楠等人[19]通过对源端语料的翻译获得目标端的关系抽取数据. 相似地, Yu 等人[8]通过回译技术来获得已有语料的复述表达. 自训练作为常见的语料扩充技术, 因其仅依赖少量样本的特点, 被广泛地应用于各式各样的分类任务中[9,20]. 本文针对低资源场景下的关系抽取任务, 在自训练的基础上, 提出复述增强和模糊数据增强的方法以提高自训练的性能.

1.2 自训练

为了获取自动标注数据, 自训练是一个历史源远流长且广泛使用的方法[21]. 近年来, 随着深度学习的发展, 自然语言处理任务对标注数据的需求越来越大, 自训练在研究界再次变得热门. 自训练被广泛应用于神经机器翻译[22]、问答[23]以及低资源场景下的句法分析任务[24]. 针对关系抽取任务, Hu 等人[25]提出使用基于元学习的自训练方法为未标注数据生成可靠的关系类别. Xu 等人[26]借助大型语言模型首先生成大量未标注数据, 随后以软标注模式标注自动生成的数据. 针对低资源场景下的关系抽取任务, 本文同样采用自训练方法获取自动标注数据. 与前人工作不同的是, 本文提出的自训练方法充分使用低置信度数据, 而这些数据由于噪音问题在前人工作中通常是被忽视的.

1.3 基于大模型的复述生成

由于文本表达的多样性特点, 复述技术是常见的数据增强方法[27,28]. 近年来, 随着大语言模型的兴起, 探索大语言模型的文本生成能力成为热点[29]. Brown 等人[30]通过实验分析了大语言模型的零样本生成能力, 即仅仅依靠目标任务的描述信息, 就能够为输入生成对应的输出. Kojima 等人[31]进一步探索了大语言模型在数学问题上的推理能力. Liu 等人[32]则是综合总结当前大语言模型的多种用法和微调范式. 针对自然语言生成任务的评价单一性问题, Tang 等人[33]利用大语言模型扩充答案多样性, 从而加强现有测试基准的评价能力. 大量研究都表明, 经过大规模数据预训练的大语言模型能够胜任许多文本生成任务, 如句子翻译、文本摘要和复述生成等. 因此, 本文使用大语言模型的复述能力为未标注数据构建可靠的复述支持包, 进而在自训练框架下提高教师模型的泛化能力, 使其能够从同一语义表达的不同文本表述中识别出高置信度的标签预测.

1.4 部分标注

为了利用模糊数据中的可用信息, 本文提出了面向模糊数据的部分标注模式. 作为一项单标签多类别任务, 针对关系抽取任务的部分标注指提供一个候选的答案集合[34]. 这一点不同于序列标注任务[35]和多标签多类别任务[36]中的部分标注方式, 前者是指仅提供序列中部分节点的标签, 后者是指仅给出答案标签集合中的部分标签. 为了训练基于部分标注的数据, 前人的工作提出了一系列研究方法[37]. Feng 等人[38]提出在自我指导式的迭代训练中融入部分标注数据的学习. 此外, Yan 等人[39]同样提出重复计算候选集中标签置信度的方法来利用部分标注数据. Wu 等人[40]针对部分标注数据的训练, 提出了一种面向候选标签集的一致性约束方法. 本文结合自训练框架下模糊数据的特性, 提出了基于负标签集合的负向训练方法.

2 背景知识

2.1 关系抽取任务定义

关系抽取任务的目标是在给定句子中识别出实体之间的语义关系[41]. 根据具体应用场景, 预定义关系集合有不同的设定. 比如, 在 SemEval-2010 Task 8 任务中[4], 关系集合主要包含因果 (cause-effect) 关系、部分-整体 (component-whole) 关系等; 而在 TACRED 任务[5]中, 关系集合主要关注面向人物和组织机构的关系, 比如夫妻关系 (per:spouse)、组织机构的别名关系 (org:alternate_names) 等. 尽管不同的关系抽取任务有着不同的需求, 但在有监督学习框架下的主体方法是通用的. 形式上, 假设标注数据集合为 $D=\{x,y\}$ 和关系集合为 R, 其中 y 表示关系类别且 $y\in R$, 而标注实例 $x=\{s,h,t\}$ 分别包含句子 s、头实体 h 和尾实体 t. 若关系分类模型为 f, 那么, 对于任意输入 x, 其目标是输出任一关系 y 的预测概率, 即 $p(y|f(x))$. 最后计算每个关系的概率, 把概率最高的关系 y^* 作为输出:

$$y^*=\arg\max_{y\in R} p(y|f(x)) \tag{1}$$

2.2 自训练

自训练通常分成以下 4 个步骤.

(1) 利用少量人工标注数据作为种子数据, 训练一个教师模型.

(2) 使用教师模型对大量无标注数据进行标签预测, 从而得到自动标注数据.

(3) 从大量自动标注数据中挑选出高置信度数据, 并以最高预测概率对应的类别作为标签构成可靠数据.

(4) 将挑选出的可靠数据与种子数据合并成新的训练集, 训练一个学生模型.

在以上步骤中, 可靠数据的选择是自训练方法有效性的关键. 本文沿用前人的工作[9,11], 当教师模型为每个未标注数据完成预测, 得到所有关系的概率分布后, 通过事先设定的概率阈值, 将预测的概率分布中最高概率大于阈值的数据作为可靠数据, 其对应的关系类别即为标签.

3 ST-LRE

如图 3 所示, 本文提出一种自训练方法 ST-LRE. 为了在低资源场景下充分利用低置信度数据, ST-LRE 包含基于复述增强的预测方法和面向模糊数据的负向训练方法. 本节首先介绍教师模型的训练方法. 随后, 给出复述数据的生成方法. 接着, 详细描述如何对自动标注数据进行分类和标签生成. 最后, 我们介绍支持可靠数据和模糊数据学习的学生模型联合训练方法.

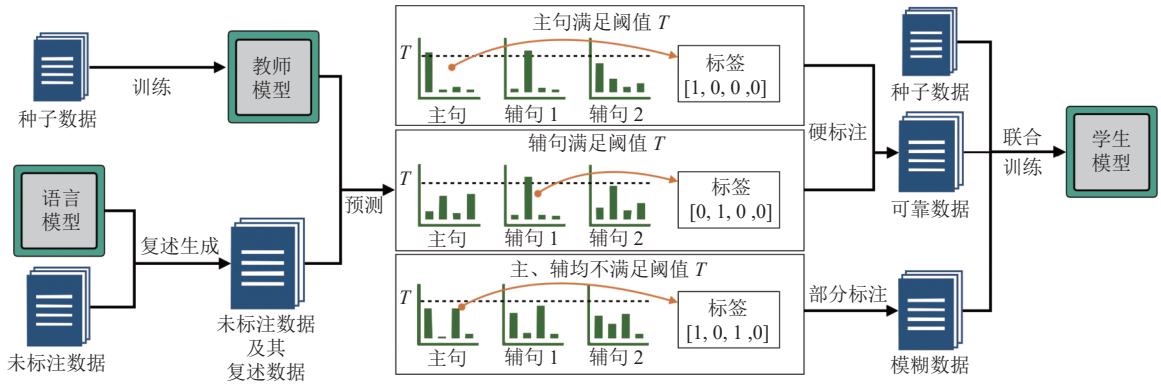


图 3 ST-LRE 自训练框架

3.1 教师模型的训练

基于人工标注的种子数据 D_{seed} 训练有监督关系抽取的教师模型 M_{teacher} 时 (如图 3 的左上部分), 本文沿用

Soares 等人[3]的“预训练模型-微调”方法: 以文本中插入实体标记符的序列作为输入, 以预训练模型作为编码层, 并添加一个分类层作为模型的输出部件. 本文采用 BERT 作为预训练模型[2]. 形式上, 对于输入实例 $x=\{s,h,t\}$, 在输入模型前, 首先被构造成如下形式:

$$x=[\mathrm{CLS}]w_0...[\mathrm{E1}]w_i...w_j[/\mathrm{E1}]...[\mathrm{E2}]w_k...w_l[/\mathrm{E2}]...w_n[\mathrm{SEP}] \tag{2}$$

其中, $w_0,w_1,...,w_n$ 是句子 s 中的每个词; 头实体 $h=\{w_i...w_k\}$; 尾实体 $t=\{w_k...w_l\}$; [CLS]、[SEP] 是 BERT 输入格式的固有字符; [E1], [/E1] 和 [E2], [/E2] 分别是用于标记头、尾实体的特殊字符. 假定以 f_{enc} 表示预训练模型 BERT 为基础的关系抽取系统中的编码层. 那么, 当输入文本进入编码层后, 我们将两个实体标记符号 ([E1] 和 [E2]) 所在位置的隐层表示提取出来, 并且拼接后作为实例 x 的关系语义表示:

$$\boldsymbol{h}=f_{\mathrm{enc}}(x)=\boldsymbol{h}_{[\mathrm{E1}]}\oplus\boldsymbol{h}_{[\mathrm{E2}]} \tag{3}$$

随后, 该关系表示被输入给分类层 f_{cls}, 用于获取每个关系类别的预测概率:

$$p(x)=f_{\mathrm{cls}}(\boldsymbol{h})=Softmax(\boldsymbol{W}\boldsymbol{h}+b) \tag{4}$$

其中, W 和 b 是分类层的模型参数.

在训练过程中, 每个输入实例 x 都带有唯一确定的标签 y, 其以 one-hot 的形式记录: 只有答案类别对应的位置记为 1, 其余都为 0 的向量. 模型参数的更新采用基于交叉熵的正向训练 (positive training) 方式:

$$L_{PT}(p(x),y)=-\sum_{i=1}^{M}y_i\log(p_i) \tag{5}$$

其中, M 为关系类别总数.

3.2 复述数据的生成

本文提出引入复述技术来提高教师模型对未标注数据的预测能力. 为此, 我们首先为未标注数据生成相应的复述数据. 近年来, 随着大语言模型的发展, 基于大语言模型的数据生成能力也逐渐增强. 得益于大规模数据的预训练和一定规模人类平行语料的微调, 大语言模型能够在给定提示指令下完成诸如文本翻译、句子复述等文本生成任务. 本文使用大语言模型的指令提示方法 (prompt) 为每一个未标注句子生成若干复述表达. 最后, 原句和复述句构成句子包的形式输入到教师模型进行自动标注.

然而, 不同于一般文本复述生成, 面向关系抽取任务的复述表达需要保证句子中给定实体对 (头、尾实体) 不能发生变化. 换句话说, 生成的复述句子中需确保原句中两个实体词的存在. 通过统计不同生成指令下生成文本中头、尾实体的情况, 我们对生成指令不断改进. 最终, 本文精心设计了如下复述生成指令来引导大语言模型生成符合关系抽取任务需求的复述数据:

Below is a sentence containing words <head > and <tail>, rewrite the input sentence in English in a different expression while keeping the words <head> and <tail>.

Input Sentence is: <input_sen>

Output Sentence is:

其中, <head> 和<tail> 是具体输入句子<input_sen> 中的头实体和尾实体. 本文采纳国内近期开源的 ChatGLM2-6B 模型[42,43], 超参数沿用其默认设置: *temperature*=0.95, *top_p*=0.7. 在实验中, 我们发现在指令的结尾处添加“keeping the words <head> and <tail>”能够有效地引导模型保留头、尾实体词. 此外, 添加“in English”来指定语言也能够减少其他语言文本的生成, 特别是针对中英文预训练的 ChatGLM2 模型.

为了保证复述句子的多样性, 我们在至多 20 次生成操作内为每个句子配备多个复述句子作为辅助句, 并与主句组成句子包: {主句,辅句$_1$,…,辅句$_N$}, 其中 N 至多为 8. 基于上述设计的实体词强调指令. 同时, 我们额外添加了基于实体词的后过滤方法来确保生成数据的质量.

3.3 数据分类与标签生成

如算法 1 所示, 本文自训练方法 ST-LRE 将自动标注数据分成 3 大类: (1) 原句或者复述句中最大预测概率大

于概率阈值的数据归类为可靠数据; (2) 至多前 K 个预测概率总和大于概率阈值的数据归类为模糊数据; (3) 剩余的则是噪音数据.

算法 1. 自动标注数据的分类.

输入: 自动标注数据 $D_{\text{auto}} = \{x, P, Y, Para(P', Y')\}$, 其中 x, P, Y 分别为主句、预测的概率分布及其关系类别; $Para(P', Y')$ 是 x 的复述数据, 其中 P' 和 Y' 分别为预测的概率分布及其关系类别;
超参数: 可靠数据的概率阈值 T, 模糊数据的候选类别数约束 K;
输出: 可靠数据 D_{con}, 模糊数据 D_{amb} 和噪音数据 D_{noisy}.

1. **for** $(x, P, Y, Para(P', Y')) \in D_{\text{auto}}$ **do**
2. $p, y \leftarrow \max(P, Y)$ // 得到主句中预测最高的概率值 p 及其对应类别 y
3. **if** $p > T$
4. Append (x, y) to D_{con}
5. **continue**
6. **else**
7. $p, y \leftarrow \max(Para(P', Y'))$ // 得到复述包中预测最高的概率值 p 及其对应类别 y
8. **if** $p > T$ **then**
9. Append (x, y) to D_{con}
10. **continue**
11. **end if**
12. **end if**
13. $P, Y \leftarrow Sort(P, Y)$ // 按照 P 中的概率值从大到小排序
14. Let $score = 0.0$
15. Let $C^+ = \varnothing$
16. **for** $(p, y) \in (P, Y)$ **do**
17. $score \leftarrow score + p$
18. Append y to C^+
19. **if** $score > T$ **then**
20. **break**
21. **end if**
22. **end for**
23. **if** $len(C^+) \leqslant K$ **then**
24. Append (x, C^+) to D_{amb}
25. **else**
26. Append x to D_{noisy}
27. **end if**
28. **end for**
29. **return** $D_{\text{con}}, D_{\text{amb}}, D_{\text{noisy}}$

3.3.1 可靠数据的筛选与标注

借助复述数据的特性, 即尽管复述句有着不同于原句的短语或结构表达方式, 但其保留了原句中的关系表达. 因此, 我们提出一个以原句为主、复述句为辅的可靠数据筛选方法. 在教师模型对未标注数据的预测阶段, 未标注数据以包含主句和辅句们的句子包形式作为输入, 并输出句子包中所有句子的概率预测分布. 算法 1 中第 2–5 行

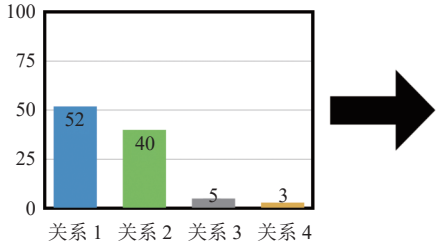
表示, 若主句的最大预测概率满足可靠数据的概率阈值, 则直接将主句 x 及其预测为最大概率的类别作为标签 y 添加到可靠数据中; 否则, 如算法 1 中第 6–12 行所示, 我们将去辅句中寻找最高的预测概率, 若其满足设定的概率阈值, 则将主句的文本内容及辅句中最大预测概率对应的类别绑定后添加到可靠数据中. 若主句和辅句都无法满足可靠数据的概率阈值条件, 则将其传至第 2 阶段的模糊数据筛选.

本文的关系抽取任务是一种单标签多分类任务, 其标签一般采用唯一的标注模式, 即 1 个句子仅属于 1 个类别. 在自训练框架下, 由于可靠数据是由高置信度预测的自动标注数据组成, 我们同样采取硬标注模式标注可靠数据, 即可靠数据的标签是一个预测时最高概率值对应的类别位置为 1, 其他位置均为 0 的 one-hot 向量.

3.3.2 模糊数据的筛选与标注

对于模糊数据的筛选, 我们提出一种贪婪的概率累加方法来获得模糊数据的候选关系类别集合, 记作正标签集合 C^+ . 如算法 1 中第 13–24 行所示, 我们首先将预测的概率分布从大到小排序, 并依次累加直到概率总和满足可靠数据的概率阈值. 同时, 将所有累加概率对应的关系类别添加到候选关系类别的正标签集合 C^+ 中. 若候选关系类别数目小于等于设定的值 K , 则将句子 x 和正标签集合 C^+ 添加到模糊数据集中, 而剩余的关系类别则被认为是负标签. 基于本文的假设“答案不在那些概率非常低的关系类别中”, 我们认为教师模型对于负标签集合中的类别作为负标签有着很高的置信度. 为了充分使用自动标注数据, 我们在实验中设置 $K = M - 1$, 其中 M 为预定义关系类别数目. 因此, 只要有 1 个类别被认为是负标签, 则该数据就被视为模糊数据, 而不满足条件的数据则被归类为不可用的噪音数据 (算法 1 中第 25–26 行). 最终, 模糊数据被标注成含有多个候选正标签的形式.

由于模糊数据的标签是包含多个候选关系类别的正标签集合, 因此, 基于唯一类别的硬标注模式并不能用于模糊数据. 否则, 由于模糊数据中最高预测概率置信度较低, 硬标注模式往往引入许多错误标注. 为了缓解噪音问题, 前人工作中提出软标注模式 (soft label) 来处理低置信度数据[9]. 在实际应用中, 通常直接使用教师模型预测的概率分布作为软标签. 在模型训练时, 以标签中每个类别位置上的值作为类别权重进行模型更新. 然而, 软标注模式并不能真正地解决噪音问题, 它只是以权重形式减少噪音的影响, 而错误标签仍然会带来错误更新. 基于上述问题, 本文提出部分标注模式来处理模糊数据. 具体来说, 基于算法 1 得到模糊数据后, 我们以正标签集合 C^+ 中的类别对应位置为 1, 其余类别对应位置为 0 的向量标注每一个模糊数据. 图 4 展示了一个面向 4 个关系类别进行分类的例子: 基于左侧教师模型预测的概率分布, 3 种不同标注方式的区别如右侧所示. 从图中可以看出, 虽然部分标注方法并不能直接告诉我们答案是哪一个 (关系 1 或关系 2), 但我们能够知道答案很大概率上不在负标签集合中 (即关系 3 和关系 4).



标注模式	关系 1	关系 2	关系 3	关系 4
硬标注	1	0	0	0
软标注	0.52	0.40	0.05	0.03
部分标注	1	1	0	0

图 4 3 种数据标注模式

3.4 学生模型的训练

当获得自动标注的可靠数据和模糊数据后, 我们将这些数据与种子数据组成新的训练集, 用于学生模型的训练. 对于可靠数据, 我们直接使用第 3.1 节中教师模型的正向训练方法. 对于模糊数据, 本文提出基于负标签集合的负向训练方法. 为了融合两种训练方式, 本文最终提出一种支持可靠数据和模糊数据的联合训练方法.

3.4.1 面向模糊数据的负向训练方法

受到 Kim 等人[44]和 Ma 等人[45]工作中关于负向训练方法 (negative training) 的启发, 不同于面向硬标注数据

的正向训练方法, 本文提出一种基于负标签集合的负向训练方法来使用模糊数据训练关系抽取系统. 形式上, 模糊数据集中每个句子都被标注一个可能的正标签集合 C^+ . 反之, 其余的类别都作为负类别聚集到负标签集合 C^- 中. 基于本文的假设: “教师模型虽然不知道答案是 C^+ 中的哪一个, 但它认为 C^- 中的负标签都不是答案”. 因此, 负向训练的核心是将模型更新定义为预测负标签概率变低的方向. 因此, 对于模糊数据中的句子 x 及其负标签集合 C^- , 我们受到 Ma 等人[45]工作的启发, 在每一轮训练前都随机从 C^- 中选取一个标签作为负标签 $\bar{y}$, 将其作为该轮训练时的标签. 因此, 最终基于交叉熵的损失函数为:

$$L_{NT}(p(x),\bar{y}) = -\sum_{i=1}^{M}\bar{y}_i \log(1-p_i) \tag{6}$$

3.4.2 支持可靠数据和模糊数据的联合训练方法

为了同时支持可靠数据和模糊数据训练, 本文提出一种支持正向和负向训练的联合训练方法. 为此, 首先引入一个标记变量 z 来表示当前输入句子是否为部分标注的模糊数据:

$$z = \begin{cases} 1, & \text{if partially labeled} \\ 0, & \text{others} \end{cases} \tag{7}$$

随后, 统一训练框架下的损失函数如下所示:

$$L(p(x),y) = -\sum_{i=1}^{M} y_i \log(|z-p_i|) \tag{8}$$

其中, $|\cdot|$ 表示取绝对值, 标签 y 对于可靠数据或种子数据来说, 是唯一的正类别位置为 1 其余为 0 的向量. 对于部分标注的模糊数据而言, 标签 y 则是负标签集合中一个随机类别对应位置为 1 而剩余为 0 的向量.

4 实 验

4.1 实验数据和设置

本文选取关系抽取任务中两个常用数据集进行实验.

• Re-TACRED: 针对 TACRED[5]中存在错误标注和冗余关系体系的问题, 由 Stoica 等人[46]修正的数据集版本. 总共包含 91 467 个标注数据, 共覆盖 39 个关系类别和 1 个特殊的其他类别.

• SemEval-2010 Task 8 (简称 SemEval): 关系抽取任务的经典数据集, 包含面向 19 个类别的 10 717 个人工标注数据. 19 个类别中包括 9 个需指出头、尾实体间关系指向的类别 (因此共 18 个) 和 1 个特殊的其他类别[4].

为了验证本文所提 ST-LRE 在低资源场景下的性能表现, 本文针对上述两个数据集进行如下处理: 种子数据由每个关系类别随机从原始人工标注训练集中选取 10 个样例组成, 而剩余数据作为未标注数据, 即保留原始句子和实体词信息, 去掉人工标注的关系类别标签. 为了进一步模拟低资源场景, 开发集规模也被限制为每个关系类别有 10 个随机样例. 同时, 特殊关系“其他类别”在两个数据集中占比很高, 比如, 在 Re-TACRED 中占比高达 66%. 为了减少不平衡问题的干扰, 本文剔除了该特殊类别. 最终, 低资源场景下两数据集的统计信息如表 1 所示.

表 1 低资源场景的数据统计

数据集	关系数	训练句子数	开发句子数	测试句子数	未标注句子数
Re-TACRED	39	390	390	5 648	18 938
SemEval	18	180	180	2 263	5 039

对于关系抽取系统的训练, 本文沿用前人工作中[3,8,26]的“预训练模型-微调”范式, 并采用 $\text{BERT}_{\text{base}}$[2]作为预训练模型. 在训练过程中, 一些超参数设置如表 2 所示, 具体数值由开发集性能决定. 本文共汇报两个评价指标, 分别是衡量全局整体性能的预测准确率 (*Accuracy* (%)) 和平衡每个关系类别性能的宏平均 $F1$ 值 (Macro-$F1$ (%)). 为了保证实验稳定性, 本文采用 5 个随机种子进行实验, 并以平均值结果汇报各个系统的结果.

表 2 超参数设置

种子数	批处理尺寸	学习率	训练轮次	提前停止轮次	概率阈值
5	16, 32	3e–5, 5e–5	20	5	0.95, 0.9, 0.85, 0.8

4.2 对比系统

为了比较本文方法 ST-LRE 与前人工作的性能差异, 本文复现了以下系统.

• Supervised: 基于有监督学习方法的系统[3], 也是本文自训练方法中的教师模型. 在训练过程中, 该方法仅使用种子数据集作为训练集.

• Confidence-ST: 基于高置信度数据的自训练方法[11]. 通过预先设定的概率阈值, 选择最大预测概率大于阈值的句子作为可靠数据, 并以硬标注模式得到标签, 其余数据则丢弃.

• HardLabel-ST: 作为 Confidence-ST 的直接比较系统, 该系统直接以硬标注模式处理所有自动标注数据, 即无论预测概率的大小, 都选用最大概率对应的类别作为标签[21].

• SoftLabel-ST: 另一种使用所有自动标注数据的方法. 不同于 HardLabel-ST 中选用最大预测概率的类别作为唯一标签的形式, 该方法把教师模型预测的概率分布作为软标签[9].

• STAD: 本文复现 Yu 等人[47]的工作, 将所有低置信度数据均转为模糊数据.

• LLM-FewShot-Random: 基于大语言模型的上下文学习 (in-context learning, ICL) 方法直接预测关系类别. 方法实现和指令设计参照 Wan 等人[48]的工作. 在实验中, 我们分别进行了 5-shot 和 10-shot 的实验, 即从种子数据集中随机选取 5 个和 10 个样例构成上下文学习的样例. 此外, 我们使用两个近期开源的大语言模型作为基石: ChatGLM2-6B[42,43]和 LLaMa2-7B-chat[49]. 经查阅公开资料, 上述两个大语言模型在预训练时均未使用本文实验涉及的 Re-TACRED 和 SemEval 数据集.

• LLM-FewShot-Retriever: 本文复现 Wan 等人[48]的工作, 在选取上下文学习的样本时, 以教师模型作为编码层提供句子的关系表示, 从而为每个测试句子寻找种子数据中与之最相似的句子. 同样地, 我们分别进行了 5-shot 和 10-shot 的实验. 大语言模型的选择同 LLM-FewShot-Random.

为了公平比较各个系统, 除基于大语言模型的系统外, 本文所提 ST-LRE 与其余比较系统在实验过程中均使用基于 $\text{BERT}_{\text{base}}$[2]的微调模型作为关系抽取模型[3], 不同之处在于数据选择策略、标注方式以及模型训练方法.

4.3 主要实验

为了验证在 Confidence-ST 系统基础上加入本文所提基于复述增强的预测方法和面向模糊数据的负向训练方法的作用, 我们在开发集上调试相关超参数, 其开发集结果如表 3 所示. 从表中可以看出, 基于复述增强 (+ paraphrase data) 的预测方法和面向模糊数据 (+ ambiguous data) 的负向训练方法在两个数据集上都取得了正面影响. 基于这两个实验结果, 最终本文提出融合复述数据和模糊数据的 ST-LRE 系统. 从表中最后一行可以看出, 本文提出的融合方法在开发集上能够取得进一步的性能提升.

表 3 复述数据和模糊数据的性能影响 (开发集)

Method	Re-TACRED		SemEval	
	Accuracy	Macro-*F*1	*Accuracy*	Macro-*F*1
Confidence-ST	75.3	74.2	78.6	74.2
+ paraphrase data	76.4	75.6	83.9	78.1
+ ambiguous data	77.2	75.8	84.1	77.2
ST-LRE	78.4	76.1	84.7	79.5

为了比较不同系统与本文所提 ST-LRE 方法在两个数据集上最终的性能表现, 我们将开发集上确定的模型应用于测试集. 实验结果如表 4 所示.

表 4 本文系统与比较系统的实验结果 (测试集)

Method		Training data			Re-TACRED		SemEval	
		D_{seed}	D_{con}	D_{amb}	*Accuracy*	Macro-*F*1	*Accuracy*	Macro-*F*1
LLM-FewShot-Random	ChatGLM2-6B 5-shot	√	×	×	17.3	18.8	15.4	10.8
	ChatGLM2-6B 10-shot	√	×	×	17.8	18.3	15.0	11.9
	LLaMa2-7B-chat 5-shot	√	×	×	32.3	41.5	21.7	17.1
	LLaMa2-7B-chat 10-shot	√	×	×	33.8	40.3	22.4	18.3
LLM-FewShot-Retriever	ChatGLM2-6B 5-shot	√	×	×	41.4	43.9	44.1	40.9
	ChatGLM2-6B 10-shot	√	×	×	45.9	46.8	50.5	47.5
	LLaMa2-7B-chat 5-shot	√	×	×	60.1	60.5	49.0	44.7
	LLaMa2-7B-chat 10-shot	√	×	×	62.9	60.9	61.2	57.2
Supervised		√	×	×	61.6	53.6	66.8	59.9
Confidence-ST		√	√	×	66.9	58.8	77.1	69.4
HardLabel-ST		√	√	√	65.3	57.3	69.6	62.5
SoftLabel-ST		√	√	√	64.2	56.6	69.2	62.6
STAD		√	√	√	74.6	62.5	78.1	72.0
ST-LRE (ours)		√	√	√	**75.8**	**63.5**	**79.5**	**73.3**

从表 4 中我们可以得出以下结论.

• 基于大语言模型的少样本上下文学习 (ICL) 方法 (LLM-FewShot-Random 或 LLM-FewShot-Retriever) 表现一般. 特别是基于随机样本的 ICL 方法, 更多的样例并没有带来性能的提升, 其性能在两个数据集上表现远低于基于 BERT 微调的 Supervised 系统. 当调用 Supervised 系统中编码层模块用于 ICL 方法的样本检索后, 性能显著提升, 特别是使用更多上下文样例时 (10-shot vs. 5-shot), 具有检索功能的 ICL 方法能够进一步取得提升. 尽管如此, 大部分结果仍低于仅使用种子数据集直接训练的 Supervised 系统, 仅有基于 LLaMa2-7B-chat 的系统在 Re-TACRED 数据集上高于 Supervised 系统. 这些实验结果表明在关系抽取任务上, 基于大语言模型的少样本 ICL 方法与基于较小规模预训练模型 (如 BERT) 的微调方法相比仍有差距.

• 相比于仅使用种子数据集 (D_{seed}) 的 Supervised 系统, 添加可靠数据 (D_{con}) 的自训练方法 Confidence-ST 取得显著提升: 在 Re-TACRED 上, *Accuracy* 值提高了 5.3% (66.9% vs. 61.6%), Macro-*F*1 值提高了 5.2% (58.8% vs. 53.6%); 在 SemEval 上, *Accuracy* 值提高了 10.3% (77.1% vs. 66.8%), Macro-*F*1 值提高了 9.5% (69.4% vs. 59.9%). 这个结果表明, 基于概率阈值筛选可靠数据的自训练方法在低资源场景下是一个简单而有效的方法.

• 相比于 Confidence-ST 系统, 额外加入模糊数据 (D_{amb}) 的 HardLabel-ST 和 SoftLabel-ST 两个自训练系统都表现不佳, 特别是在 SemEval 数据集上, HardLabel-ST 和 SoftLabel-ST 的 *Accuracy* 值下降分别达到了 7.5% 和 7.9%, Macro-*F*1 值下降分别达到了 6.9% 和 6.8%, 降幅明显. 尽管两者性能均优于 Supervised 系统, 但我们认为这主要源于可靠数据带来的性能提升. 这些结果说明在自训练框架下, 无论是使用最高概率的硬标注模式还是使用概率分布的软标注模式, 当引入低置信度数据时不可避免地会带入大量噪音, 导致性能下降.

• 本文所提 ST-LRE 不仅相比于 Supervised 系统取得了显著的提升 (在两个评价指标上性能提升都在 10% 以上), 而且 ST-LRE 的性能表现显著优于使用相同文本数据 ($D_{\text{con}} + D_{\text{amb}}$) 但不同标注和训练方式的 HardLabel-ST 和 SoftLabel-ST 两个系统. 相比于 Confidence-ST 系统, 本文方法能够取得明显的提升: 在 Re-TACRED 上, *Accuracy* 值提升达到了 8.9% (75.8% vs. 66.9%), Macro-*F*1 值提升达到了 4.7% (63.5% vs. 58.8%); 在 SemEval 上, *Accuracy* 值提高了 2.4% (79.5% vs. 77.1%), Macro-*F*1 值提高了 3.9% (73.3% vs. 69.4%). 此外, 相比于不使用复述增强的 STAD 系统, 本文的 ST-LRE 能够取得进一步的提升. 这说明本文所提复述增强方法能够在抑制低置信度数据中噪音的同时, 充分利用其包含的有用信息.

5 实验分析

5.1 消融实验

ST-LRE 不仅引入了复述数据帮助教师模型筛选可靠数据, 而且提出了基于低置信度数据的模糊数据学习方法. 因此, 我们针对复述增强 (paraphrase data) 的数据预测方法以及模糊数据使用和训练中涉及的部分标注方法和负向训练方法分别进行消融实验.

实验结果如表 5 所示. 从表 5 中可以看出: (1) 如果没有引入复述数据 (– paraphrase data), 实验结果均有 1% 以上的下降. 由此可以看出, 在低资源场景的自训练方法基础上引入复述数据是能够帮助筛选更多可靠数据的. (2) 对于模糊数据的使用, 若不使用部分标注方法 (– partial label), 即以硬标注模式处理模糊数据但依旧使用负向训练方法, 则性能有明显的降低. 以 *Accuracy* 值为例, 在 Re-TACRED 和 SemEval 上性能分别下降 8.6% 和 2.2%. 该方法的性能下降是可以预见的: 硬标注模式在很多情况下会导致正确类别出现在负标签集合中, 因此, 在负向训练过程中有可能选择到正确类别作为负标签, 导致模型往错误方向更新. (3) 若去掉负向训练 (– negative training), 即以部分标注模式处理模糊数据, 但采用正向训练方法, 也就是训练时标签随机选自正标签集合, 根据表中实验结果, 这种策略导致性能急剧下降. 同样以 *Accuracy* 为指标, 在两个数据集上的性能下降分别达到 24.4% 和 6.9%. 其主要原因是由于正标签集合中至多有 1 个正确标签 (甚至全是错误标签), 从而正向训练时极有可能选取到错误标签对模型进行正向更新. 因此, 性能大幅度下降也是符合直觉的. 由此可见, 针对模糊数据的使用, 本文提出的部分标注模式和负向训练方法缺一不可, 同时也证明本文针对模糊数据的假设在实验应用中是可行的.

表 5 消融实验结果 (测试集)

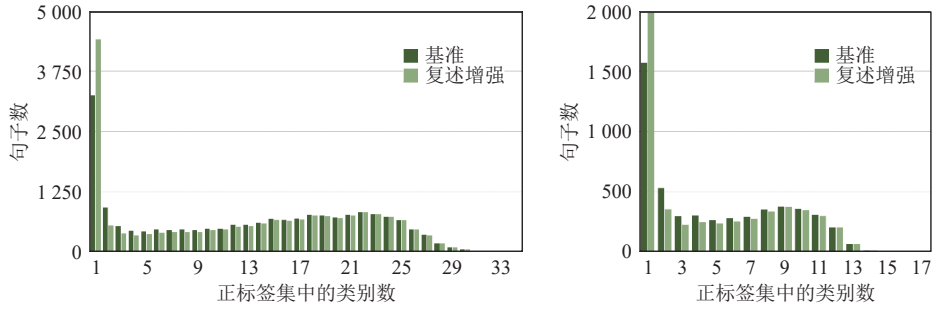
Method	Re-TACRED		SemEval	
	Accuracy	Macro-*F*1	*Accuracy*	Macro-*F*1
ST-LRE	75.8	63.5	79.5	73.3
– paraphrase data	−1.2	−1.0	−1.4	−1.3
– partial label	−8.6	−4.1	−2.2	−1.7
– negative training	−24.4	−13.7	−6.9	−6.1

5.2 自动标注数据的分布

为了分析自训练框架下自动标注数据的分布情况, 我们统计 Re-TACRED 和 SemEval 两个数据集中教师模型预测数据的分布情况. 数据分布如图 5 所示, 其中纵坐标表示句子数目, 而横坐标表示正标签集中的数目, 即为满足概率阈值条件 (此处概率阈值为 0.9), 以概率累加方式构成的正标签集中所包含的类别数目. 当横坐标值为 1 时表示正标签集合中只有 1 个类别, 即可靠数据, 而其余大于 1 的情况为模糊数据. 作为比较, 我们分别分析教师模型 (详见第 3.1 节) 直接预测数据为基准的分布和引入复述数据 (详见第 3.3.1 节) 进行复述增强预测的分布.

从图 5 中可以看出, 一方面, 可靠数据的数量占比较大. 在 Re-TACRED 和 SemEval 中, 可靠数据的句子数分别达到了 3 258 和 1 568 句, 分别占全量的 17.2% 和 31.1%. 模糊数据虽然单独来看每个数量都不多, 但总量多于可靠数据. 在本文实验中, 由于两个关系抽取任务的预定义关系类别数均比较大以及概率阈值的原因, 我们在两个数据集中都没有发现纯噪音数据. 因此, Re-TACRED 和 SemEval 上分别有 15 680 和 3 471 句模糊数据, 占比分别达到 82.8% 和 68.9%. 这表明忽视如此大量的模糊数据是不合适的.

另一方面, 当引入复述增强方法后, 可靠数据明显增多: 在 Re-TACRED 和 SemEval 上可靠数据分别提升至 4 419 和 1 992 句, 占比提升至 23.3% 和 39.5%. 在第 4.3 节中的实验结果表明, 基于复述增强的预测方法能够带来性能的提升, 而性能的提升显然来自新增的可靠数据. 同时, 我们还发现大部分新增可靠数据来源于正标签集合中类别数相对较少的集合. 当正标签集合中类别数目较大时, 即令教师模型感到非常困惑的句子, 基于复述增强的预测方法也很难再将模糊数据识别为可靠数据.



(a) Re-TACRED (b) SemEval

图 5 自动标注数据分布情况

5.3 样例分析

图 6 展示了 Re-TACRED 中 3 个典型复述增强预测的例子, 其中每个例子包括一个原句和一个复述句子. 左侧展示原始英文句和对应中文翻译, 并以“$[]_{e1}$”和“$[]_{e2}$”标识出头、尾实体词, 右侧则是教师模型预测原句和复述句分别为正确关系类别的概率. 第 1 个例子中原句与复述句的主要不同在于, “was in business from”被替换为“was established in”. 相比于前者, 后者显示地表达“公司被建立”的含义, 因此, 教师模型将其预测为正确关系“成立日期”的概率从 37% 提升至 92%. 在第 2 个例子中, 原句中的头尾实体相距甚远, 且使用了后置定语“his youngest daughter”的形式来表达头尾实体间的关系; 复述句中头尾实体紧密相连且使用简单结构“Mona Kempfer, Herry's youngest daughter”来体现两实体之间的关系. 因此, 教师模型预测其为“子女关系”的概率从 40% 提升至 93%. 第 3 个例子是句式结构的变化, 将用于表达出生含义的“was born in”短语直接连接头尾实体, 使得预测其为“出生城市”的概率从 41% 提升至 94%. 最终, 若设定概率阈值为 0.9, 则 3 个句子受益于其复述句子的预测结果, 都能够被选为可靠数据.

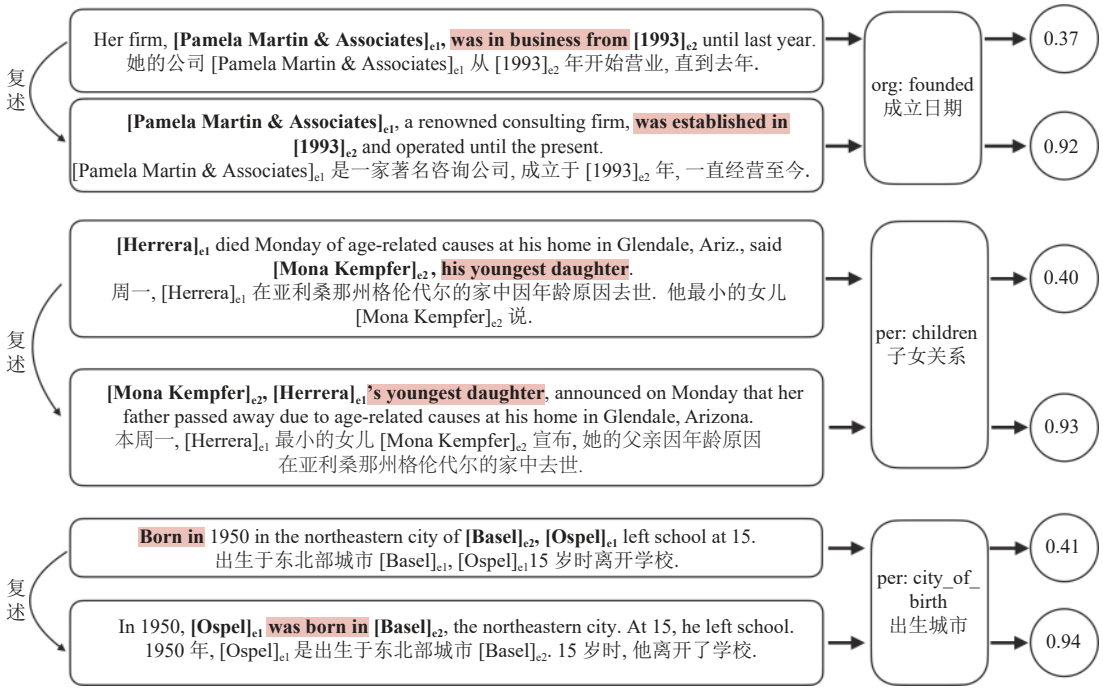


图 6 复述增强预测样例

图 7 展示了 Re-TACRED 中 3 个典型模糊数据的样例, 其中左侧表示英文原句和对应中文翻译, 右侧列出了按预测概率排序的前几个关系类别及其概率值. 第 1 个例子中由于同时提到了“was born in”和“grew up in”两种表达, 教师模型对“出生城市”和“居住城市”两个关系类别都预测了较高的概率. 在第 2 个例子中, 由于缺少头尾实体的类型信息, 仅基于常用于表达附属关系的“X's Y”短语结构使得教师模型对“机构所在国家、机构所在城市和机构隶属于”3 种关系都预测了较高的概率. 对于第 3 个例子, 虽然看起来很简单: Richard Lindzen 的头衔是 Professor, 但是, 由于短语“MIT Professor Richard Lindzen”不仅受到 Richard Lindzen 与 MIT 之间是“就职于”关系的影响, 而且鉴于 MIT 的学校属性, 它经常会与人物构成“就读学校”的关系. 因此, 教师模型最终无法给出确切的预测. 在上述 3 个例子中, 最高预测概率对应的关系类别并非正确答案, 而答案分别是各自预测概率第 2 高的关系类别. 因此, 在自训练框架下, 若无视阈值, 直接选取最高概率对应的类别作为标签 (硬标注模式), 则将引入 3 个错误数据. 若设定较高的概率阈值来筛选可靠数据 (比如 0.9), 则这 3 个句子都会被弃用. 显然上述两种方法都无法利用这类数据. 但在本文基于概率累加的部分标注模式下, 图 7 中所示例子都会被添加到模糊数据集中, 对应标签则是满足条件的前几个关系类别组成的候选标签集合, 而剩余关系类别则均被认为是负标签. 最终, 基于负向训练方法可以充分利用这 3 个句子中的负标签信息来训练关系抽取系统.

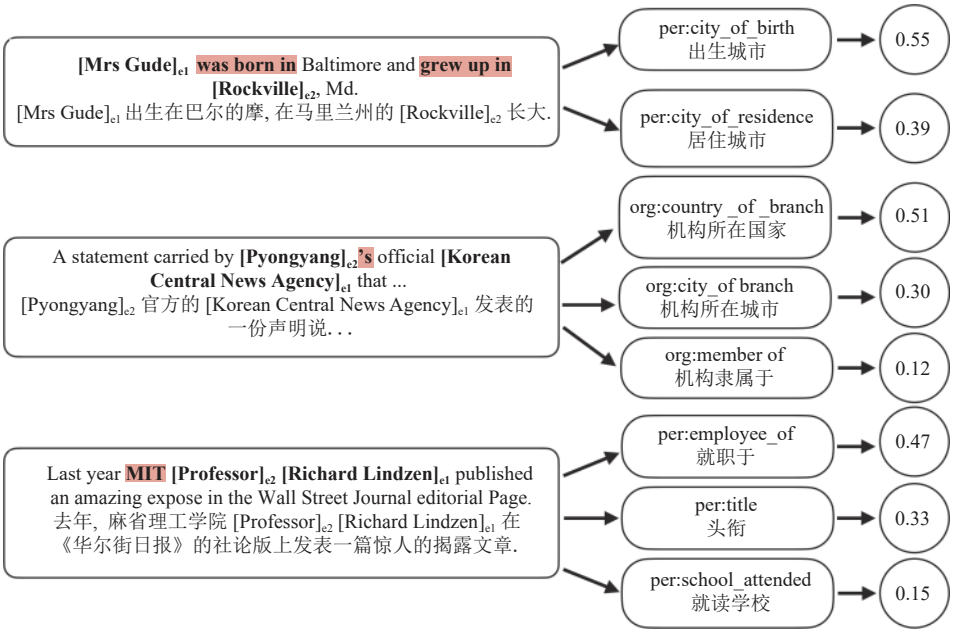


图 7　模糊数据样例

6 总　结

针对自训练方法在低资源关系抽取任务中无法充分利用大量低置信度数据的问题, 本文提出一种有效利用低置信度数据的方法来改进自训练. 该方法首先基于大语言模型的复述生成能力扩充未标注数据的表达多样性, 进而提出基于复述增强的预测方法来加强可靠数据的筛选能力. 其次, 针对大量难以利用的低置信度数据, 提出基于概率累加的部分标注方法将其转化为可利用的模糊数据, 进而提出基于负标签集合的负向训练方法, 使用模糊数据训练关系抽取系统. 最终, 关系抽取系统的训练融合了可靠数据和模糊数据. 在两个关系抽取任务上的实验结果表明, 本文方法能够有效解决自训练方法在低资源场景下教师模型泛化能力差和低置信度数据利用率低的问题, 并最终提升关系抽取系统的性能.

References:

[1] Zhou GD, Su J, Zhang J, Zhang M. Exploring various knowledge in relation extraction. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor: ACL, 2005. 427–434. [doi: 10.3115/1219840.1219893]

[2] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[3] Soares LB, Fitzgerald N, Ling J, Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 2895–2905. [doi: 10.18653/v1/P19-1279]

[4] Hendrickx I, Kim SN, Kozareva Z, Nakov P, Séaghdha DÓ, Padó S, Pennacchiotti M, Romano L, Szpakowicz S. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proc. of the 5th Int'l Workshop on Semantic Evaluation. Uppsala: ACL, 2010. 33–38.

[5] Zhang, YH, Zhong V, Chen DQ, Angeli G, Manning CD. Position-aware attention and supervised data improve slot filling. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 35–45. [doi: 10.18653/v1/D17-1004]

[6] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Suntec: ACL, 2009. 1003–1011. [doi: 10.5555/1690219.1690287]

[7] Luo F, Nagesh A, Sharp R, Surdeanu M. Semi-supervised teacher-student architecture for relation extraction. In: Proc. of the 3rd Workshop on Structured Prediction for NLP. Minneapolis: ACL, 2019. 29–37. [doi: 10.18653/v1/W19-1505]

[8] Yu JJ, Zhu T, Chen WL, Zhang W, Zhang M. Improving relation extraction with relational paraphrase sentences. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 1687–1698. [doi: 10.18653/v1/2020.coling-main.148]

[9] Xie QZ, Luong MT, Hovy E, Le QV. Self-training with noisy student improves ImageNet classification. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10684–10695. [doi: 10.1109/CVPR42600.2020.01070]

[10] Qian LH, Zhou GD, Kong F, Zhu QM. Semi-supervised learning for semantic relation classification using stratified sampling strategy. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2009. 1437–1445.

[11] Du JF, Grave E, Gunel B, Chaudhary V, Celebi O, Auli M, Stoyanov V, Conneau A. Self-training improves pre-training for natural language understanding. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 5408–5418. [doi: 10.18653/v1/2021.naacl-main.426]

[12] Amini MR, Feofanov V, Pauletto L, Hadjadj L, Devijver E, Maximov Y. Self-training: A survey. arXiv:2202.12040, 2022.

[13] Xu X, Chen X, Zhang NY, Xie X, Chen X, Chen HJ. Towards realistic low-resource relation extraction: A benchmark with empirical baseline study. In: Proc. of the Findings of the Association for Computational Linguistics. Abu Dhabi: ACL, 2022. 413–427. [doi: 10.18653/v1/2022.findings-emnlp.29]

[14] Ouyang DT, Qu JF, Ye YX. Extending training set in distant supervision by ontology for relation extraction. Ruan Jian Xue Bao/Journal of Software, 2014, 25(9): 2088–2101 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4638.htm [doi: 10.13328/j.cnki.jos.004638]

[15] Yang S, Zhang YF, Niu GL, Zhao QH, Pu SL. Entity concept-enhanced few-shot relation extraction. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 987–991. [doi: 10.18653/v1/2021.acl-short.124]

[16] Dong MQ, Pan CG, Luo ZP. MapRE: An effective semantic mapping approach for low-resource relation extraction. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 2694–2704. [doi: 10.18653/v1/2021.emnlp-main.212]

[17] Liu Y, Hu JP, Wan X, Chang TH. A simple yet effective relation information guided approach for few-shot relation extraction. In: Proc. of the Findings of the Association for Computational Linguistics. Dublin: ACL, 2022. 757–763. [doi: 10.18653/v1/2022.findings-acl.62]

[18] Zhu SY, Hui HT, Qian LH, Zhang M. Family relation extraction from Wikipedia by self-supervised learning. Journal of Computer Applications, 2015, 35(4): 1013–1016, 1020 (in Chinese with English abstract). [doi: 10.11772/j.issn.1001-9081.2015.04.1013]

[19] Hu YN, Shu JG, Qian LH, Zhu QM. Cross-lingual relation extraction via machine translation. Journal of Chinese Information Processing, 2013, 27(5): 191–198 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2013.05.028]

[20] Zoph B, Ghiasi G, Lin TY, Cui Y, Liu HX, Cubuk ED, Le Q. Rethinking pre-training and self-training. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 323. [doi: 10.5555/3495724.3496047]

[21] Lee DH. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proc. of the Workshop on Challenges in Representation Learning. 2013. 896.

[22] Zhang JJ, Zong CQ. Exploiting source-side monolingual data in neural machine translation. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 1535–1545. [doi: 10.18653/v1/D16-1160]

[23] Sachan M, Xing E. Self-training for jointly learning to ask and answer questions. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL, 2018. 629–640. [doi: 10.18653/v1/N18-1058]

[24] Rotman G, Reichart R. Deep contextualized self-training for low resource dependency parsing. Trans. of the Association for Computational Linguistics, 2019, 7: 695–713. [doi: 10.1162/tacl_a_00294]

[25] Hu XM, Zhang CW, Ma FK, Liu CY, Wen LJ, Philip SY. Semi-supervised relation extraction via incremental meta self-training. In: Proc. of the Findings of the Association for Computational Linguistics. Punta Cana: ACL, 2021. 487–496. [doi: 10.18653/v1/2021.findings-emnlp.44]

[26] Xu BF, Wang Q, Lyu YJ, Dai D, Zhang YD, Mao ZD. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 8186–8207. [doi: 10.18653/v1/2023.acl-long.455]

[27] Zhao SQ, Liu T, Li S. Research on paraphrasing technology. Ruan Jian Xue Bao/Journal of Software, 2009, 20(8): 2124–2137 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3587.htm [doi: 10.3724/SP.J.1001.2009.03587]

[28] Zhu HY, Jin ZL, Hong Y, Su YL, Zhang M. Directional data augmentation for question paraphrase identification. Journal of Chinese Information Processing, 2022, 36(9): 38–45 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2022.09.004]

[29] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. In: Proc. of the 10th Int'l Conf. on Learning Representations. 2022.

[30] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Conf. on Neural Information Processing Systems. 2020. 1877–1901.

[31] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proc. of the 36th Conf. on Neural Information Processing Systems. 2022. 22199–22213.

[32] Liu PF, Yuan WZ, Fu JL, Jiang ZB, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 2023, 55(9): 195. [doi: 10.1145/3560815]

[33] Tang TY, Lu HY, Jiang YE, Huang HY, Zhang DD, Zhao WX, Kocmi T, Wei FR. Not all metrics are guilty: Improving NLG evaluation by diversifying references. arXiv:2305.15067, 2024.

[34] Cour T, Sapp B, Taskar B. Learning from partial labels. The Journal of Machine Learning Research, 2011, 12: 1501–1536. [doi: 10.5555/1953048.2021049]

[35] Li ZH, Zhang M, Chen WL. Ambiguity-aware ensemble training for semi-supervised dependency parsing. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: ACL, 2014. 457–467. [doi: 10.3115/v1/P14-1043]

[36] Xie MK, Huang SJ. Partial multi-label learning with noisy label identification. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 6454–6461. [doi: 10.1609/aaai.v34i04.6117]

[37] Nguyen N, Caruana R. Classification with partial labels. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Las Vegas: ACM, 2008. 551–559. [doi: 10.1145/1401890.1401958]

[38] Feng L, An B. Partial label learning with self-guided retraining. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 3542–3549. [doi: 10.1609/aaai.v33i01.33013542]

[39] Yan Y, Guo YH. Partial label learning with batch label correction. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 6575–6582. [doi: 10.1609/aaai.v34i04.6132]

[40] Wu DD, Wang DB, Zhang ML. Revisiting consistency regularization for deep partial label learning. In: Proc. of the 39th Int'l Conf. on Machine Learning. 2022. 24212–24225.

[41] E HH, Zhang WJ, Xiao SQ, Cheng R, Hu YX, Zhou XS, Niu PQ. Survey of entity relationship extraction based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(6): 1793–1818 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5817.htm [doi: 10.13328/j.cnki.jos.005817]

[42] Zeng AH, Liu X, Du ZX, Wang ZH, Lai HY, Ding M, Yang ZY, Xu YF, Zheng WD, Xia X, Tam WL, Ma ZX, Xue YF, Zhai JD, Chen

WG, Zhang P, Dong YX, Tang J. GLM-130B: An open bilingual pre-trained model. arXiv:2210.02414, 2023.

[43] Du ZX, Qian YJ, Liu X, Ding M, Qiu JZ, Yang ZL, Tang J. GLM: General language model pretraining with autoregressive blank infilling. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 320–335. [doi: 10.18653/v1/2022.acl-long.26]

[44] Kim Y, Yim J, Yun J, Kim J. NLNL: Negative learning for noisy labels. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 101–110. [doi: 10.1109/ICCV.2019.00019]

[45] Ma RT, Gui T, Li LY, Zhang Q, Huang XJ, Zhou YQ. SENT: Sentence-level distant relation extraction via negative training. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 6201–6213. [doi: 10.18653/v1/2021.acl-long.484]

[46] Stoica G, Platanios EA, Póczos B. Re-TACRED: Addressing shortcomings of the tacred dataset. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 13843–13850. [doi: 10.1609/aaai.v35i15.17631]

[47] Yu JJ, Wang X, Zhao JJ, Yang CJ, Chen WL. STAD: Self-training with ambiguous data for low-resource relation extraction. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: Int'l Committee on Computational Linguistics, 2022. 2044–2054.

[48] Wan Z, Cheng F, Mao ZY, Liu QY, Song HY, Li JW, Kurohashi S. GPT-RE: In-context learning for relation extraction using large language models. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 3534–3547. [doi: 10.18653/v1/2023.emnlp-main.214]

[49] Touvron H, Martin L, Stone K, *et al*. LLAMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.

附中文参考文献:

[14] 欧阳丹彤, 瞿剑峰, 叶育鑫. 关系抽取中基于本体的远监督样本扩充. 软件学报, 2014, 25(9): 2088–2101.http://www.jos.org.cn/1000-9825/4638.htm [doi: 10.13328/j.cnki.jos.004638]

[18] 朱苏阳, 惠浩添, 钱龙华, 张民. 基于自监督学习的维基百科家庭关系抽取. 计算机应用, 2015, 35(4): 1013–1016, 1020. [doi: 10.11772/j.issn.1001-9081.2015.04.1013]

[19] 胡亚楠, 舒佳根, 钱龙华, 朱巧明. 基于机器翻译的跨语言关系抽取. 中文信息学报, 2013, 27(5): 191–198. [doi: 10.3969/j.issn.1003-0077.2013.05.028]

[27] 赵世奇, 刘挺, 李生. 复述技术研究. 软件学报, 2009, 20(8): 2124–2137. http://www.jos.org.cn/1000-9825/3587.htm [doi: 10.3724/SP.J.1001.2009.03587]

[28] 朱鸿雨, 金志凌, 洪宇, 苏玉兰, 张民. 面向问题复述识别的定向数据增强方法. 中文信息学报, 2022, 36(9): 38–45. [doi: 10.3969/j.issn.1003-0077.2022.09.004]

[41] 鄂海红, 张文静, 肖思琪, 程瑞, 胡莺夕, 周筱松, 牛佩晴. 深度学习实体关系抽取研究综述. 软件学报, 2019, 30(6): 1793–1818. http://www.jos.org.cn/1000-9825/5817.htm [doi: 10.13328/j.cnki.jos.005817]

郁俊杰(1992－), 男, 博士生, 主要研究领域为自然语言处理, 信息抽取.

陈文亮(1977－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 信息抽取, 知识图谱.

王星(1988－), 男, 博士, 高级研究员, 主要研究领域为自然语言处理, 机器翻译, 大语言模型.

张民(1970－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器翻译, 人工智能.