

编号 \_\_\_\_\_



南京航空航天大学

# 本科毕业设计（论文）

题 目

基于主动学习的关系抽取  
方法研究

学生姓名	蔡益武
学 号	161630324
学 院	计算机科学与技术学院
专 业	软件工程
班 级	软件工程卓越班
指导教师	黄圣君 教授

二〇二〇年六月

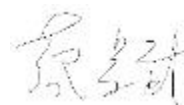


# 南京航空航天大学

## 本科毕业设计（论文）诚信承诺书

本人郑重声明：所呈交的毕业设计（论文）是本人在导师的指导下独立进行研究所取得的成果。尽我所知，除了文中特别加以标注和致谢的内容外，本设计（论文）不包含任何其他个人或集体已经发表或撰写的成果作品。对本设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

作者签名：



日期： 2020 年 5 月 26 日

# 南京航空航天大学

## 毕业设计（论文）使用授权声明

本人完全了解南京航空航天大学有关收集、保留和使用本人所送交的毕业设计（论文）的规定，即：本科生在校攻读学位期间毕业设计（论文）工作的知识产权单位属南京航空航天大学。学校有权保留并向国家有关部门或机构送交毕业设计（论文）的复印件和电子版，允许论文被查阅和借阅，可以公布论文的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编论文。保密的论文在解密后适用本声明。

论文涉密情况：

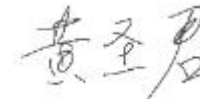
☒ 不保密 ☐ 保密

保密期（起讫日期：至）

作者签名：



导师签名：



日期： 2020 年 5 月 26 日

日期： 2020 年 5 月 28 日



## 摘 要

当代社会的知识已呈爆炸性增长，其中最常见的知识蕴含在非结构化的自然语言文本当中。信息抽取（**Information Extraction**）技术通过一组被提及的实体、这些实体之间的关系以及这些实体参与的事件来表达非结构化文本其中的语义知识。作为信息抽取中关键的一环，关系抽取（**Relation Extraction**）技术，通过判断给定实体之间所属关系，为文本知识理解提供了重要的理论依据和使用价值。

目前基于监督学习的关系抽取需要大量已标记样本，随机选择部分数据标记不仅是对数据资源的浪费，还会直接影响到分类模型最终的准确率。事实上，随着数据收集和储存技术的发展，获取大量未标记自然语言文本变得十分容易，因此设计一种能够有效利用未标记样本集进行关系抽取的算法具有重要的实际价值。

为了解决上述问题，本文以主动学习为切入点，实现了多种采样算法，主要有不确定性，多样性，代表性等算法，在验证主动学习适用于关系抽取任务的基础上，通过融合多种采样标准最终获得一个可以在多个数据集和多种学习模型下仍具有有效性的主动学习样本选择策略。

实验证明，本文提出的多标准融合采样策略是一个具有有效性、健壮性的策略，与多个单策略采样算法相比，在多个数据集上都能够取得相当或者更高的分类精度。

**关键词：** 主动学习，深度学习，关系抽取，多标准

## **ABSTRACT**

Knowledge in contemporary society has been growing explosively. The most common knowledge is contained in unstructured natural language texts. Most of its semantic knowledge can be expressed by a group of mentioned entities, the relationship between these entities and the events that these entities participate in. As a key part of information extraction, relationship extraction has important theoretical significance and practical application.

At present, a large number of labeled samples are needed in relation extraction based on supervised learning. Random annotation is not only a waste of data resources, but also affects the performance of classifier. In fact, with the development of data collection and storage technology, it is very easy to obtain a large amount of unlabeled natural language text. It is of great practical value to design an algorithm that can effectively use the unlabeled sample set.

In order to solve the above problems, the paper takes active learning as the starting point, designs and implements a variety of sample selection algorithms, including uncertainty, representativeness, diversity, etc. On the basis of verifying the applicability of active learning to relation extraction tasks, an active learning sample selection strategy which can be effective in multiple datasets and learning models is finally obtained by integrating multiple sampling standards.

Experiments show that the multi-standard fusion sampling strategy proposed in this paper is an effective and universal strategy. Compared with multiple single strategy sampling algorithms, it can achieve comparable or higher classification accuracy on multiple datasets.

**KEY WORDS:** Active Learning, Deep Learning, Relation Extraction, Multi-Criteria

# 目 录

第一章 引言 .....	1
1.1 研究背景 .....	1
1.2 国内外研究现状 .....	2
1.2.1 关系抽取研究现状 .....	2
1.2.2 主动学习研究现状 .....	3
1.3 研究目标与内容 .....	4
1.4 本文组织结构 .....	4
1.5 本章小结 .....	5
第二章 相关工作 .....	6
2.1 关系抽取 .....	6
2.2 主动学习 .....	9
2.2.1 主动学习算法模型 .....	9
2.2.2 主流主动学习介绍 .....	10
2.3 本章小结 .....	11
第三章 基于主动学习的关系抽取 .....	12
3.1 抽取模型 .....	12
3.2 基本采样方法 .....	14
3.2.1 基于不确定性的采样方法 .....	14
3.2.2 基于多样性的采样方法 .....	15
3.2.3 基于代表性的采样方法 .....	16
3.3 集成采样策略 .....	16
3.3.1 基于多标准的赋权采样策略 .....	17
3.3.2 基于多标准的逐层采样策略 .....	18
3.4 本章小结 .....	19
第四章 实验及分析 .....	20
4.1 实验环境与数据 .....	20
4.2 实验评价指标 .....	21
4.3 实验结果与分析 .....	21
4.4 本章小结 .....	25
第五章 总结与展望 .....	26
致谢 .....	27
参考文献 .....	27





## 第一章 引言

### 1.1 研究背景

随着网络技术的不断更迭，以文本、语音、图像等数据方式为基础的信息大量充斥在人们的日常生活，“信息过载”由此而生。在进行信息搜索时，人们期待计算机能真正地理解他所需要的信息并作出高质量反馈，而不仅仅是返回包含关键字的文档或部分语句，这就要求信息抽取技术能迅速、精准地返回用户所需的信息。

于是，知识图谱（Knowledge Graph）应运而生。为了提高当前的搜索引擎的反馈质量，知识图谱以结构化的形式将互联网的各类信息组织成更便于人类理解的形式，提供了一个更好的存储、查询互联网各式信息的能力。当用户进行搜索时，知识图谱先通过分析问题语义，理解用户真实查询需求，然后在图谱数据库中查询相关知识并返回给用户，从而改进搜索质量。知识图谱依托于大数据和人工智能，已经成为管理互联网知识的基础设施。

在构建知识图谱的过程中，信息抽取是最为核心的技术，包括实体抽取（Entity Extraction）、关系抽取等，其中关系抽取是信息抽取中不可或缺的一环，关系抽取通过判断所给句子实体之间是否存在某种关系，并进一步确定该关系类别，将文本分析从语言层面提升到内容层面。通常以（实体 1，关系类别，实体 2）的三元组形式表示实体 1 和实体 2 之间存在某种类别的语义关系，比如句子“CPU 是计算机的核心部件”中表示的实体关系是（计算机，核心部件，CPU）。实体之间不同的关系将独立的实体关联起来，形成知识网络，高质量的关系抽取不仅能够增大知识图谱的规模，而且能够保障知识图谱的质量，因此，探索研究关系抽取技术，具有一定的理论意义和实际的应用价值。

传统的关系抽取方法中，研究人员需要手工精心设计语义规则，根据不同样本与不同规则的匹配给出该样本中实体之间的关系，但这种方法不仅需要特定领域专家参与，而且还难以迁移到其他领域，因此代价极高。随着深度神经网络的成功，基于对数据进行表征学习的方法广泛应用于关系抽取，取代了传统的基于手工特征<sup>[2-4]</sup>、核函数<sup>[5]</sup>、条件随机场<sup>[6]</sup>的方法，代表模型有卷积神经网络（Convolutional Neural Network, CNN），循环神经网络（Recurrent Neural Network, RNN）等。然而，此类有监督学习想要取得良好的性能需要大量的标记样本，为了避免人工标记数据尤其是海量非结构化的网络数据时的费时费力问题，可以考虑使用主动学习（Active Learning）技术。

主动学习旨在通过尽可能小的标记代价训练出一个有效的学习模型。通过启发式学习策略主动选择对模型最有帮助的样本交由人类专家标记，并将标记的实例加入训练集中，

通过迭代训练以提高分类器的泛化性能。随着信息时代各类数据的指数级增加，数据的标记问题越来越受到学术界和工业界的重视，主动学习作为减轻数据标记代价的有效途径，在理论和算法上都取得了显著进展，已被广泛地应用在图片处理<sup>[2]</sup>、语音识别<sup>[2]</sup>等领域。

本文主要研究如何在关系抽取任务中应用主动学习技术，其意义是，在小规模已标记语料的条件下，能够有效利用大规模未标注语料中的潜在信息来学习以及选择最有效的部分语料进行人工标注，以保证关系抽取模型在较小的标记代价下，取得较高的学习性能。

## 1.2 国内外研究现状

### 1.2.1 关系抽取研究现状

关系抽取是文本内容理解的重要支撑技术之一，为问答系统、信息检索等应用提供大量支持，本节主要介绍限定域下面向非结构化文本的关系抽取典型技术。

传统的基于特征的方法，提取不同的特征集并将其输入到所选择的分类器（例如，逻辑回归，SVM 等），如 Rink and Harabagiu 等人<sup>[2]</sup>利用词法语法在内的各种特征，最后用 SVM 分类器对实体间的关系进行分类。这种方法严重依赖于精心设计的特征来学习良好的模型，而且由于不同训练数据集的覆盖率很低，泛化性能很差。

基于核函数的方法通过指定两条语句之间的某种相似性度量，而避免显式的构造特征表示，如 Bunescu 等人<sup>[2]</sup>根据两个实体之间的最短依赖路径设计了一个最短依存树核。但核函数方法的一个潜在困难是所有数据信息都由内核函数（相似性度量）完全概括的，但设计有效的内核往往需要专家依赖句法分析等专业知识且在该领域有多年的深耕。

近年来，基于神经网络的方法在关系抽取中占据主导地位。2014 年，Zeng 等人<sup>[2]</sup>首先将卷积神经网络（CNN）应用在有监督关系抽取任务中，他们用向量的形式表示句子内的所有词语，再利用 CNN 和最大池化得到句子的向量表示，最后通过 softmax 分类器预测两个实体之间的关系。同时期 Thien 等人<sup>[2]</sup>，Santos 等人<sup>[2]</sup>的工作也是采用了类似的方法。

然而，CNN 会损失句子中各个词汇的位置信息，不适合学习长距离的语义信息，Cai 等人<sup>[2]</sup>提出使用双向循环卷积神经网络（BRCNN），其使用长短期记忆（Long Short-term Memory, LSTM）层学习最短依赖路径（Shortest Dependency Paths, SDP）的全局信息，使用卷积提取词向量之间的局部信息，并提出了双向结构同时学习 SDP 的前后向句子表示，大大提高了对实体关系的分类性能。

2016 年，Zhou 等人<sup>[2]</sup>提出用基于注意力（Attention）机制的双向长短期记忆（Att-BLSTM）抽取实体关系。首先将语料样本输入到模型中，将该样本中每个词映射为对应的

词向量，用 BLSTM 提取词向量特征，获得句子的整体表示，再利用注意力机制<sup>[2]</sup>，使模型训练重点集中在对关系分类起更高作用的词向量上。这种基于注意力的模型可以自动聚焦于句子中最重要的语义信息，从而无需使用额外的知识和 NLP 系统即可获得良好的抽取性能。

2018 年，谷歌研究院 Devlin 等人提出语言模型 BERT<sup>[2]</sup>，全称为 Bidirectional Encoder Representations for Transformers。BERT 的特征抽取器采用 Transformer<sup>[2]</sup>，相对 RNN 能捕捉更长距离的依赖，以及拥有良好的并行能力，对比之前的预训练模型，如 GPT（Generative Pre-Training）等，它捕捉到的是真正意义上的双向上下文信息，大部分自然语言任务包括关系抽取都能够通过 BERT 微调得到很好的效果。如，Wu S 等人<sup>[2]</sup>提出的 R-BERT，就是以 BERT 为基础的关系抽取模型。通过结合来自目标实体的信息，将句子和实体对信息合并为输入序列，经过 BERT 中的 Transformer 提取生成序列的深层语言表示向量，拼接句子向量和实体向量得到综合向量，对其做非线性和线性变化，最终通过 softmax 分类得到关系类别。

### 1.2.2 主动学习研究现状

由于机器学习尤其是有监督学习严重依赖大量标记数据对分类器进行迭代训练，所以实际落地的场景很少。在这种情况下，主动学习应运而生并且发展迅速，主动学习（Active Learning）方法研究最早始于 20 世纪 90 年代，主要解决机器学习任务中所面临的样本标记代价高的问题<sup>[2]</sup>，不同于一般的监督学习，主动学习起初只有少量已标记样本，迭代查询部分未标记样本，获取其真实标记，并更新已标记样本集重新训练分类器，直到分类器达到一定精度或超过目标标记代价。

主动学习的核心思想是，从大量的未标记样本中循环选择对模型帮助最大的样本人工标记并循环训练以获得较高的分类模型性能，从而以更少的标记代价获得更好的抽取效果。因此，主动学习是一种可以适用于文本中实体关系分类的算法。

根据可供选择的未标注样本集的特性，主动学习算法可以分为查询合成（query synthesis）算法、基于流（stream-based）算法以及基于池（pool-based）算法<sup>[2]</sup>。其中，基于池场景的主动学习因为其符合大量实际的应用场景，得到广泛研究和实际应用。池场景下常用的主动选择样本方法有不确定性采样、泛化误差减小、最小化版本空间和密度加权方法等。

近年来，Sun 等人<sup>[2]</sup>使用关注词组相似性的全局分类器和关注语法特征的局部分类器来发现分类面附近的未标记样本，通过仅标记这些样本以达到用更少的标记代价获得更好

的抽取效果。Angeli 等人<sup>[2]</sup>提出一种主动学习算法，精心挑选少量具有不确定性且具有代表性的实例，这些实例为远程监督的关系抽取提供部分监督，显著提升了算法性能。Huang 等人<sup>[2]</sup>提出一种结合未标记实例的信息性和代表性，同时引入标签之间的相关性的系统性算法，该方法在单标签和多标签学习中都有着突出的表现。Hsu 等人<sup>[2]</sup>设计了一种“混合”学习算法可以使得机器根据不同数据集使用不同的主动学习算法，借此提高主动学习算法的迁移性。一般的主动学习策略一次迭代只挑选一个实例供专家标记，模型在已标记数据集更新之后重新训练，这样的做法十分耗时，Lourentzou 等人<sup>[2]</sup>通过实验表明使用适当的批次大小可以使得模型性能在特定领域几乎不变的前提下，减少训练时间。

主动学习目前已经在多个任务中得到了实际应用。Zhu 等人<sup>[2]</sup>以 SVM 作为基准模型，使用样本到分类超平面的距离作为评价标准对短文本分类，同样的，Almgren 等人<sup>[2]</sup>以类似的方法做入侵检测分析；医学图像领域中，Yakoub 等人<sup>[2]</sup>通过深度置信网络学习图像特征，用 DNN 神经网络作为基准分类器，采用最大信息量策略对心电图进行医学上的分析；Yao 等人<sup>[2]</sup>以卷积神经网络作为基准分类器，使用最大信息量的样本选择策略对工业图像做图像检测。

### 1.3 研究目标与内容

本工作的主要目标是设计一种基于主动学习的关系抽取方法，得到一个具有高普适性和可迁移性的样本查询策略，在不同抽取模型、不同数据集上与随机采样对比，验证其有效性。

本文的主要研究内容包括：

1. 实现多种主动学习算法，包括基于不确定性、代表性、多样性的基本采样方法，验证主动学习算法在关系抽取任务中的有效性。
2. 在基本采样方法的基础上，设计融合多种标准的采样策略，最后在关系抽取任务上进行对比实验，得到最优采样策略。
3. 实现多种关系抽取模型，包括 CNN、BLSTM、R-BERT 等，验证设计的主动学习算法具有普适性。

### 1.4 本文组织结构

针对当前研究现状存在的问题以及文本的研究目标、研究内容，拟定本文的基本结构如下：

一、引言。本章首先指出研究基于主动学习的关系抽取方法的背景和意义，概括性地

介绍了国内外当前关系抽取和主动学习的研究现状，在此基础上，确定了本文研究目标和内容，最后给出全文的组织结构。

二、相关工作。介绍关系抽取的任务中常用的分类模型和基本算法，以及描述主动学习的算法流程、分类情况。

三、基于主动学习的关系抽取。具体介绍本文提出的基于主动学习的关系抽取方法，在分析了基于不确定性、多样性、代表性三类基本的样本采样策略的基础上，提出多标准融合采样策略，进一步提高主动学习的提升性能。

四、实验及分析。验证提出的基于主动学习的关系抽取方法的有效性、健壮性、可迁移性。

## 1.5 本章小结

本章主要介绍关系抽取和主动学习领域的研究背景、研究现状以及指出目前存在的问题，在此基础上，明确本次毕业设计的研究意义，给出解决方法，最后简要介绍本文的组织结构。

## 第二章 相关工作

### 2.1 关系抽取

关系抽取作为知识图谱构建的一环，针对非结构化的文本提取其中实体关系三元组，组成结构化的知识。传统的基于人工设计规则的关系抽取工作，严重依赖人工设计的特征及抽取特征的质量，有很大的局限性，而随着深度学习在图像、语音等领域取得的成功，许多关系抽取的研究工作也引入神经网络以自动提取句子中的特征，不仅减少了特征工程所需精力，而且可以取得良好的抽取效果。

基于深度学习的抽取模型将关系抽取当作是一个多分类问题，模型框架如图 2.1所示：

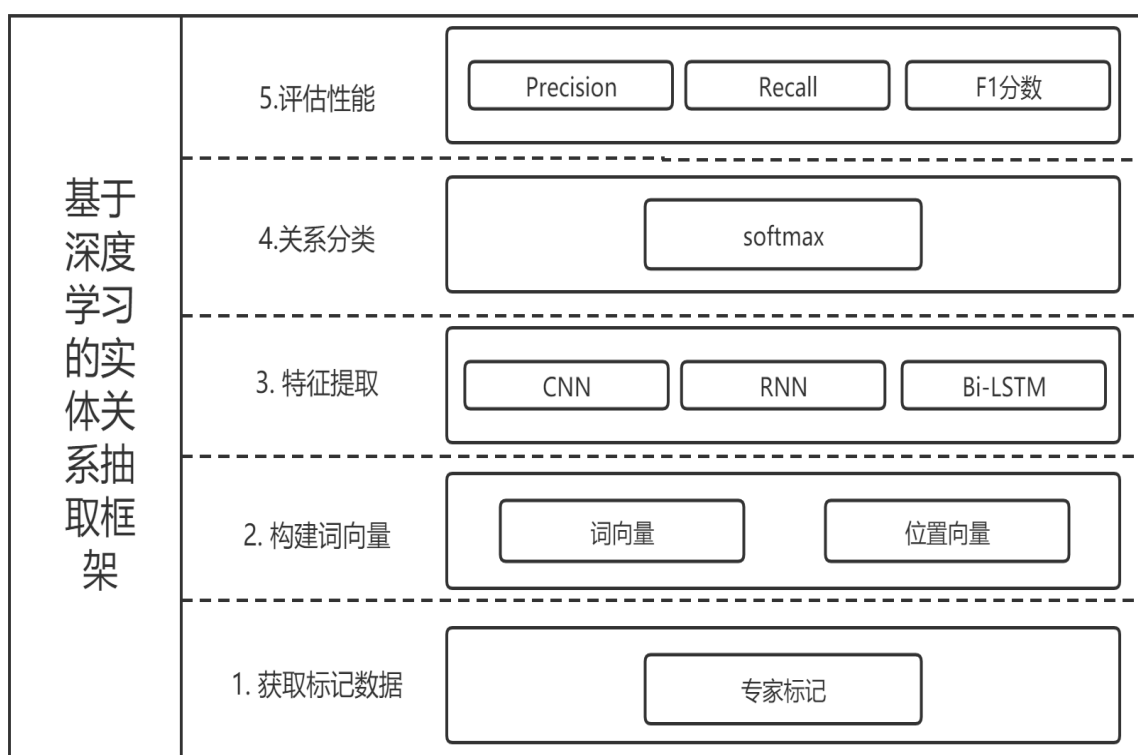


图 2.1 基于深度学习的实体关系抽取框架

1. 获取标记数据：通过专家标记获得带标签数据集。
2. 构建词向量：通过对已标记文本分词，并将其映射成对应的词向量。
3. 特征提取：将由词向量组成的句向量送入有监督的分类器提取句子特征。
4. 关系分类：句子特征向量经过线性/非线性变化之后，送入 softmax 分类，得到目标实体关系。
5. 评估性能：通过 F1 分数等指标对关系分类结果进行评估。

使用卷积神经网络进行关系抽取时，模型首先通过预训练或者随机初始化的词嵌入（Word Embedding）将句子中的词表示为词向量，通过拼接句子中的实体词向量及其上下文相对位置表征实体词的位置向量得到最终的词向量表示，随后通过 CNN 网络抽取句子级别的特征  $C_1, C_2, \dots, C_n$ ，接着通过池化层和全连接层得到句子的关系类别。

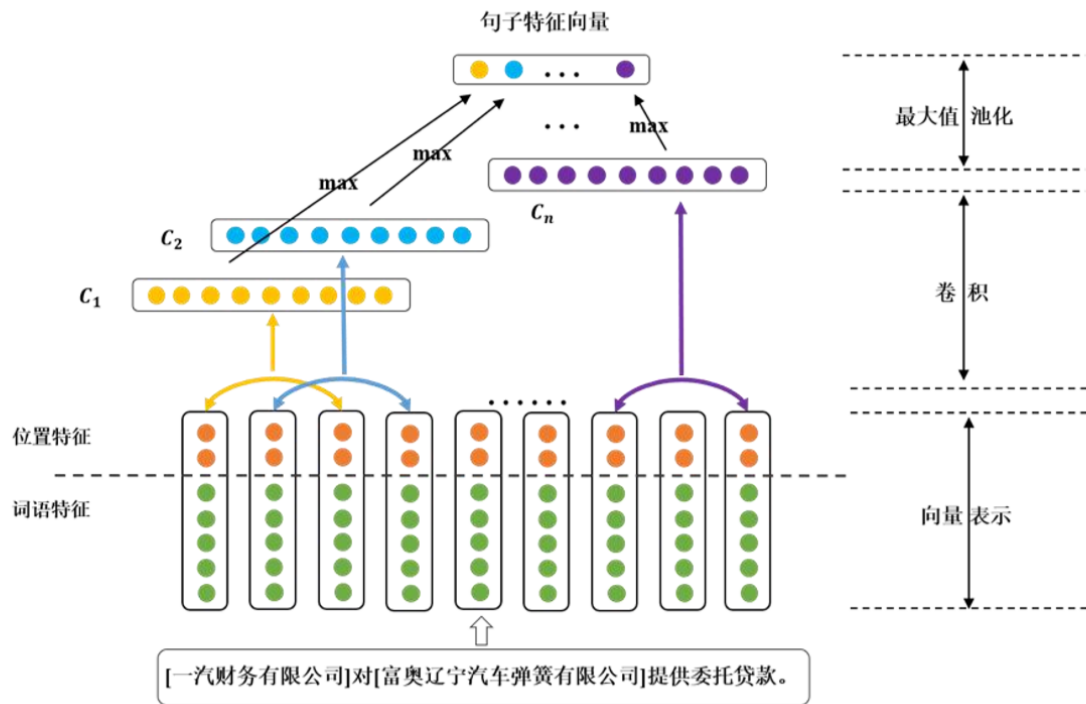
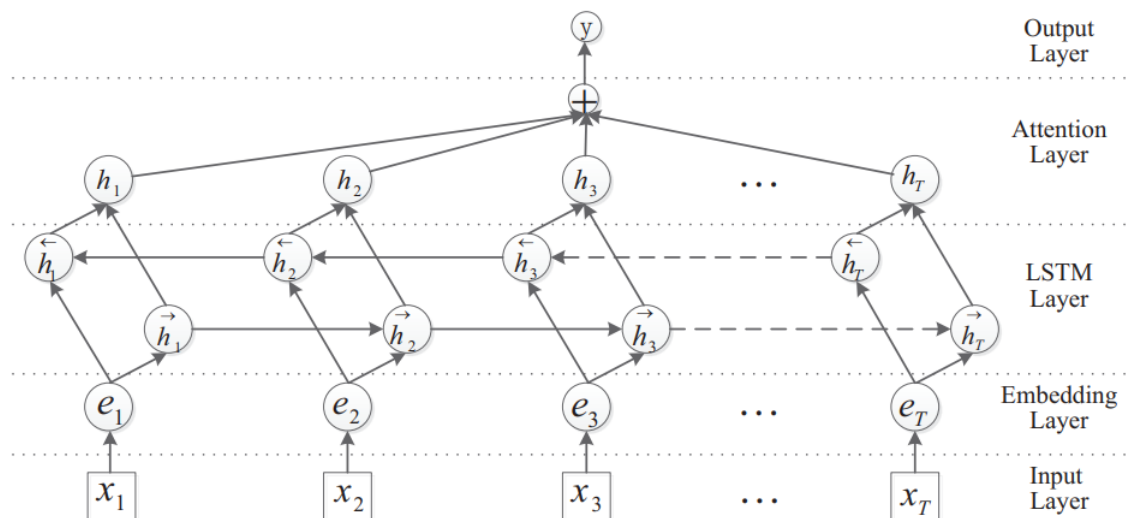
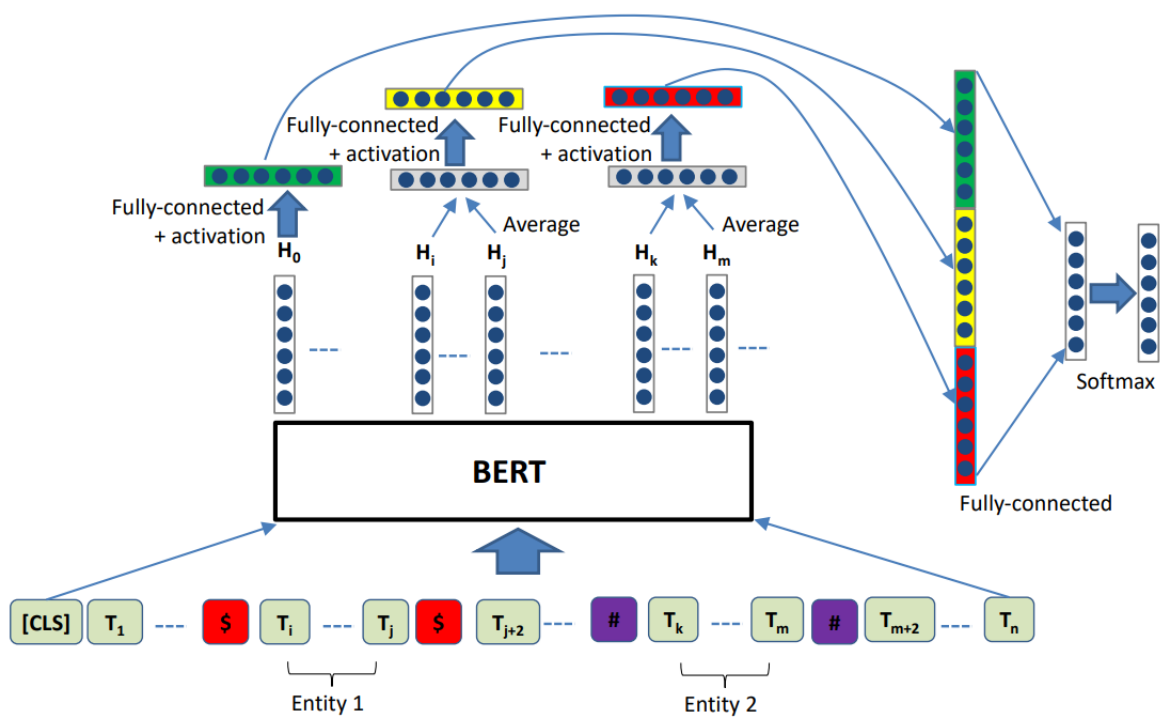


图 2.2 基于 CNN 的关系抽取模型

基于注意力机制的 BLSTM 神经网络模型开展关系抽取任务，首先通过 Embedding 层，将每个词映射到低维空间，紧接着双向 LSTM 从中获取高级特征，然后与在注意力层生成的权重向量相乘，使每一次迭代中的词汇级的特征合并为句子级的特征，最终将句子级的特征向量用于关系分类。模型结构如图 2.3。

BERT 是一种以 Transformer 为特征编码器的预训练双向语言模型，而 Transformer 是基于自注意力机制叠加而成的深度网络，不仅能够捕获长距离特征，而且有良好的并行计算能力，使用 R-BERT 模型处理中文关系任务，利用预训练的 BERT 语言模型，结合来自目标实体的信息，根据 BERT 的输入要求在实体前后添加标识符的方式表明实体位置，将输入句子和实体对信息合并为输入序列，BERT 会输出标识符最终隐含状态向量和两个目标实体的最终隐含状态向量，综合三部分的向量信息经过线性/非线性变化，最后通过 softmax 层进行分类。模型结构如图 2.4所示。

图 2.3 基于 Att-BLSTM 的关系抽取模型<sup>[2]</sup>图 2.4 基于 BERT 的关系抽取模型<sup>[2]</sup>



## 2.2 主动学习

### 2.2.1 主动学习算法模型

主动学习通过设计合理的采样方法，从未标记样本中选择出最能帮助基础模型获得更好性能的样本进行标记后，加入已标记训练集，重新训练基础模型，反复迭代，直到模型性能达到一定要求或者超出标记代价。在整个主动学习过程中，最核心的就是学习引擎和采样引擎<sup>[2]</sup>。

学习引擎指的是基础模型，即分类器在已标记数据集上进行循环训练，并在测试集上验证泛化性能。而采样引擎指的是通过使用实例选择算法在未标记数据集上选择部分待标记样本，这些样本经过人类专家标记之后，供学习引擎使用。整个学习过程就是学习引擎和采样引擎不断迭代工作，最终在标记代价可接受的条件下得到一个性能足够优秀的分类器。图 2.5 为主动学习模型图。

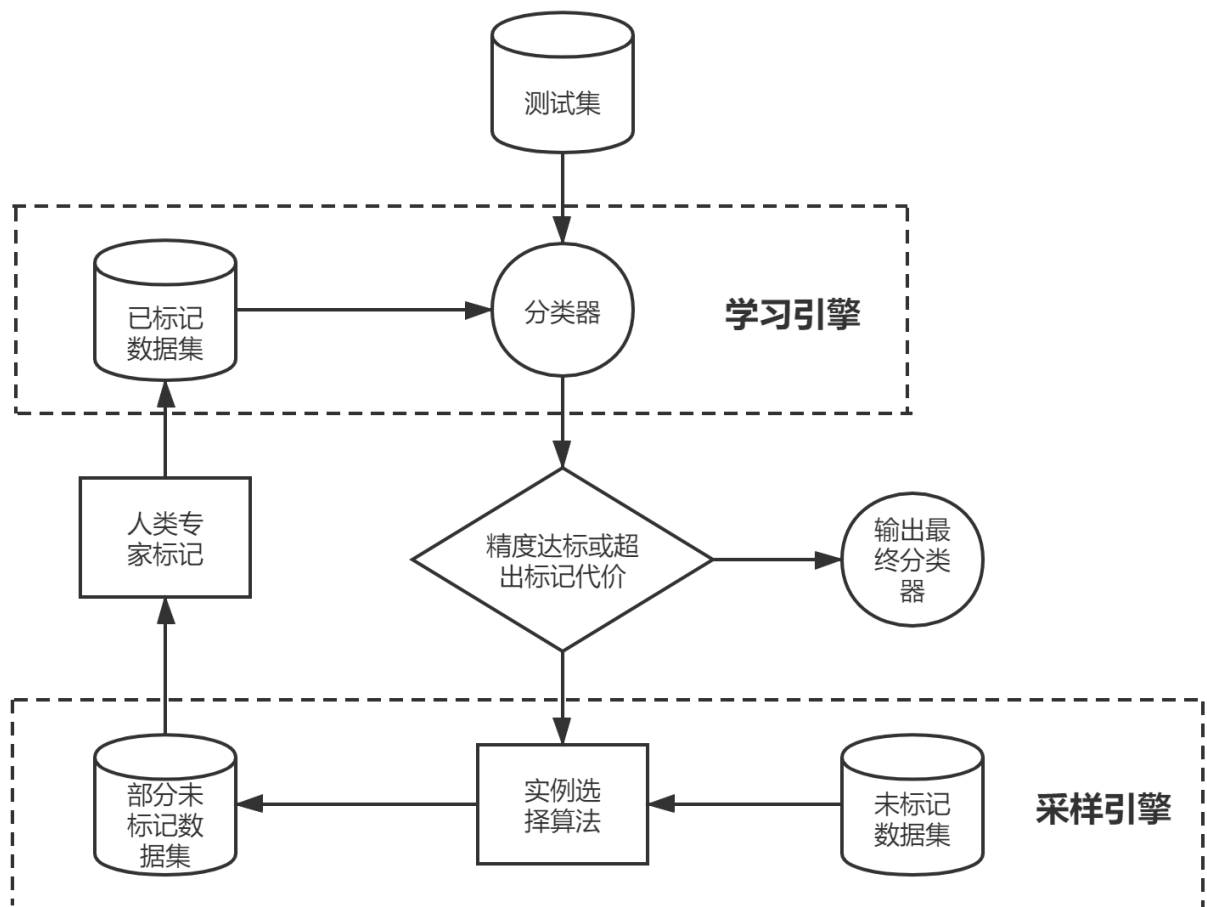


图 2.5 主动学习模型图

### 2.2.2 主流主动学习介绍

根据不同应用场景下主动学习挑选未标记样本的方式不同，将主动学习算法分为以下三种<sup>[2]</sup>：

- 查询合成（query synthesis）算法
- 基于流（stream-based）算法
- 基于池（pool-based）算法

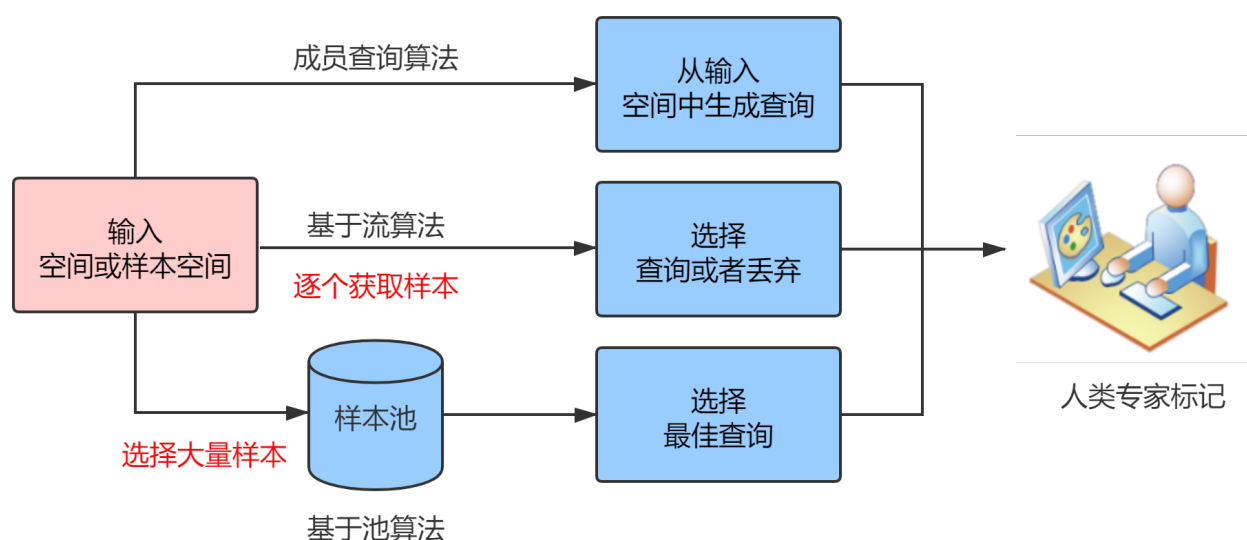


图 2.6 三种主动学习算法

查询合成算法是第一种通过查询样本进行学习的算法<sup>[2]</sup>，通过向专家提问的方式获取整个输入空间中对学习效果最有帮助的样本标记。查询合成算法在限定域下是十分有效的，但是，该方法查询输入空间中所有的样本而不考虑样本的实际分布会导致产生大量合法却没有实际意义的工作<sup>[2]</sup>，例如，当样本数据是文本形式，基础模型是生成式对抗模型（Generative Adversarial Networks, GAN）时，该算法可能会创建与正常语句词法相似但没有实际语义信息的文本，将这些数据样本交由专家标注显然是没有意义的。因此，此类算法不适合由人工专家标记样本的应用场景。

为了解决上述问题，研究人员提出了基于流的采样策略，这类做法将落在样例空间中的所有未标记样本按顺序根据采样策略决定标记或者丢弃<sup>[2]</sup>。一般而言，这种采样策略需要逐个将未标记样本的信息含量跟事先设定好的固定阈值做比较，因此，无法得到未标记样本的整体结构分布及样本间的差异。仅适用于入侵检测<sup>[2]</sup>、信息获取<sup>[2]</sup>等场景。

针对这一缺陷，研究人员提出基于池的采样策略<sup>[2]</sup>。将所有未标记样本视为一个“池”，

从样本池中有选择性地标记样本，与基于流的算法相比，其通过计算样本池中所有未标记样本的信息含量并从中挑选出信息含量最好的样本进行标记，避免了设定固定阈值，查询无意义样本的情况，因而成为主动学习领域中研究最广泛的一类算法，在视频检索<sup>[2][1]</sup>、文本分类<sup>[2][1]</sup>、信息抽取<sup>[2][1]</sup>等领域都有具体的应用。

## 2.3 本章小结

本章简要介绍关系抽取任务中三种主流的抽取模型，本文将采用 R-BERT 作为关系抽取基础模型。着重介绍了主动学习算法模型，当前主流的三种主动学习算法，得出基于池的采样策略更适合关系抽取情景。

## 第三章 基于主动学习的关系抽取

### 3.1 抽取模型

想要得到实体之间的关系类别，往往需要有效结合目标实体和句子语义信息的关系，抽取模型 R-BERT 利用 BERT 强大的编码能力，同时抽取文本中的两种特征，进行关系分类。BERT 分为预训练和微调两个阶段，预训练阶段以 Transformer 为特征编码器，通过海量语料预训练，联合执行两个任务——掩蔽语言模型（MLM）和下一句预测（NSP），得到序列的局部和全局特征表示，BERT 的网络结构如图 3.1 所示。

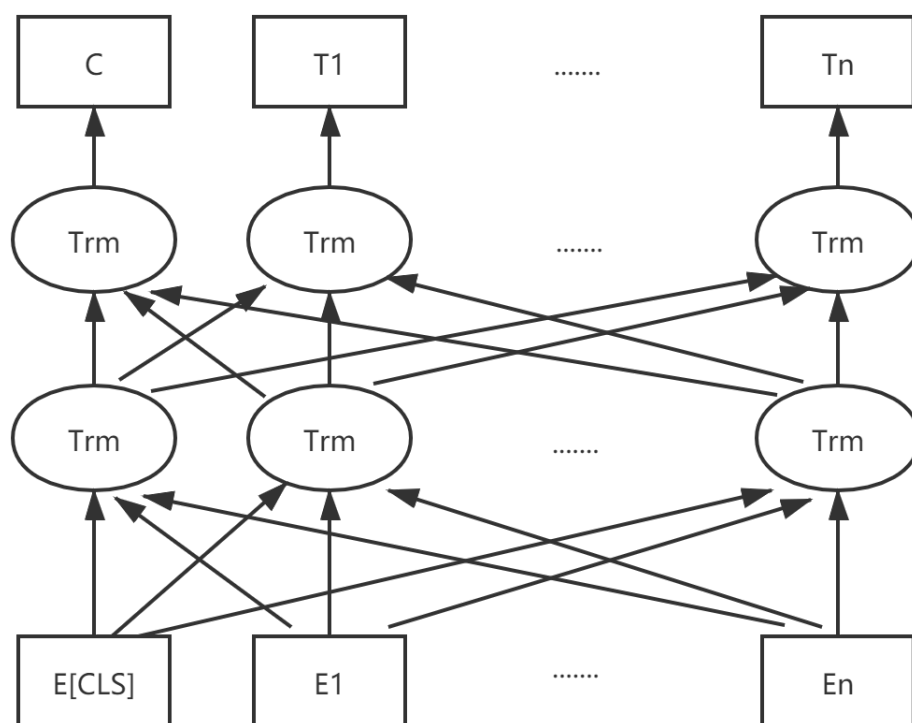


图 3.1 BERT 模型结构

不难看出，BERT 本质上是一堆堆叠在一起的 Transformer 编码器（图中的 Trm 模块）。Transformer 起源于注意力（Attention）机制，其完全抛弃了传统的 RNN，整个网络结构完全是由 Attention 机制组成。通过 Transformer 编码器提取特征，不仅拥有良好的并行计算能力能够快速训练而且能够捕捉语句间的深层联系，是目前十分流行的特征提取器。

BERT 的输入向量由词向量、位置向量、段向量三部分组成，以及，为了便于后续微调阶段执行分类任务，在每个输入序列的开头加上 [CLS] token。在执行 MLM 任务时，通过随机遮住 15% 的 token 作为训练样本，在这些样本中，80% 用 mask token 替代，10% 用

随机的一个 token 替代，10% 保持这个 token 不变，再让编码器对这些 token 根据上下文做预测，在这样的任务驱动下，通过迭代训练，BERT 模型可以很好地学习到每一个 token 的语法、句法、上下文特征。

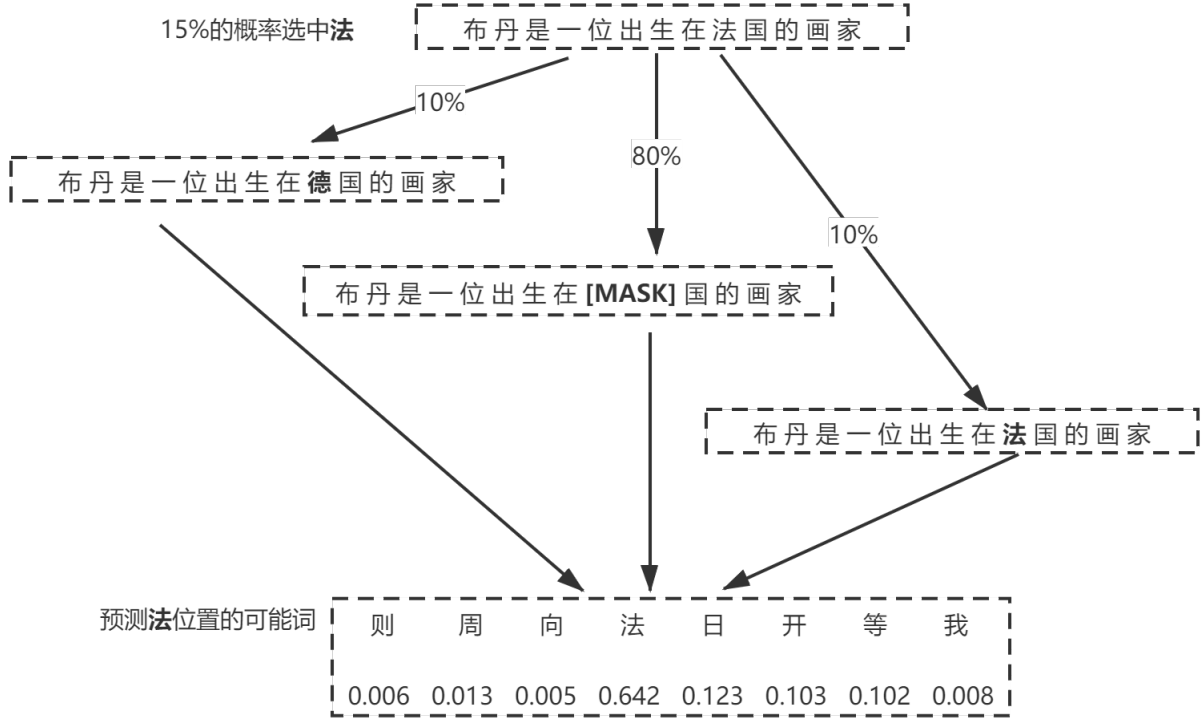


图 3.2 MLM 任务执行过程

由于在预训练阶段，[CLS] token 不参与掩蔽，因而，该位置面向整个序列的所有位置做 Attention，使得 [CLS] 位置的输出足够表达整个句子的信息，而其他 token 对应的向量更关注该 token 的语义语法及上下文信息表达，至此，经过预训练，得到一个泛化性能足够优秀的能够表征语料信息的语言模型。

在微调阶段，针对关系抽取任务，为了使 R-BERT 能够定位两个实体的位置，在第一个前后添加特殊字符"\$"，在第二个实体前后添加特殊字符"#”，利用三部分特征进行最后的关系分类，包括 [CLS] 最终隐含状态向量，两个实体的隐含状态向量。具体过程如下：

把 [CLS] 位置的输出向量记为  $H_0$ ，实体  $e_1$  的各个字向量是  $H_i$  到  $H_k$ ，实体  $e_2$  的各个字向量是  $H_j$  到  $H_m$ ，两个实体的输出向量表示为：

$$H_1 = \frac{1}{k-i+1} \sum_{t=i}^k H_t \quad (3.1)$$

$$H_2 = \frac{1}{m-j+1} \sum_{t=j}^m H_t \quad (3.2)$$

其中  $H_1, H_2 \in R_{n \times d}$ ,  $n$  是批量大小,  $d$  是 BERT 的隐藏状态大小。两个实体向量和 [CLS] 位置输入向量经过非线性激活 ( $\tanh$ ) 再经由全连接层得到  $H'_0, H'_1, H'_2$ 。

$$H'_0 = W_0[\tanh(H_0)] + b_0 \quad (3.3)$$

$$H'_1 = W_1[\tanh(H_1)] + b_1 \quad (3.4)$$

$$H'_2 = W_2[\tanh(H_2)] + b_2 \quad (3.5)$$

其中  $b'_{0-2} \in R_{d \times d}$ ,  $W'_{0-2} \in R_{d \times d}$ ,  $H'_{0-2} \in R_{n \times d}$ ,  $W_1 = W_2$ ,  $b_1 = b_2$ 。即  $W_1$  和  $W_2$ ,  $b_1$  和  $b_2$  共享参数。通过拼接这三个特征向量并输入到全连接层, 最后使用 softmax 分类得到关系类别。

$$h' = W_3[\text{concat}(H'_0, H'_1, H'_2)] + b_3 \quad (3.6)$$

$$p = \text{softmax}(h') \quad (3.7)$$

其中,  $W_3 \in R_{l \times 3d}$ ,  $h' \in R_{n \times l}$ ,  $l$  是关系种类数,  $p$  是预测的关系类别。

## 3.2 基本采样方法

本文主要以基于池的样本采样策略为研究对象, 研究制定适合关系抽取任务的样本采样策略, 在保证模型达到一定性能的同时尽可能降低标注成本, 即维护一个未标记样本池, 通过主动学习样本策略迭代选择样本标记后训练模型, 使得模型的泛化能力得到快速提升, 其选择策略一般遵循贪心思想, 即每次迭代从未标记样本集中选择某一属性最大 (或最小) 的样本进行标记。

### 3.2.1 基于不确定性的采样方法

基于不确定性的样本采样方法的选择策略的主要思想是从未标注样本集中选择分类模型给出较低置信度的样本, 通过比较模型分类结果中的各样本的置信度大小, 判断各待选样本能给分类器带来的信息含量, 从未标注样本集中选择能带来最大信息量的样本获取标注加入标注样本集。具体到实体关系抽取任务中, 由于一条语句有多种可能的标记, 可以用关系被预测为每一种类别的置信度来衡量样本的不确定性。least confident 算法以每一个样本可能性最大的分类作为其代表分类, 并根据代表类别对应的置信度从中选择不确定性最高 (即置信度最低) 的样本:

$$x_L^* = \arg \max_{x \in U} 1 - P_\theta(\hat{y}|x) \quad (3.8)$$

其中

$$\hat{y} = \arg \max_y P_\theta(y|x) \quad (3.9)$$

但是以上方法只考虑了后验概率最高的那个类，简单忽略了其他类，很可能会出现度量误差。实际分类结果中，经常出现置信度的最高两个类别预测概率接近的情况，针对这种情况对上面的最小置信度策略做改进，margin sampling 采用每个样本的预测结果中最大和次大的类别置信度的差作为样本不确定性的衡量标准，显然置信度差值越小的样本的实际类别更难区分，因此选择该类样本获取标注能够给基础模型带来更多的有效信息，采样策略如式 3.10 所示：

$$x_M^* = \arg \min_{x \in U} \{P_\theta(\hat{y}_f|x) - P_\theta(\hat{y}_s|x)\} \quad (3.10)$$

其中

$$\hat{y}_f = \arg \max_y P_\theta(y|x) \quad (3.11)$$

$$\hat{y}_s = \arg \max_{y, y \neq \hat{y}_f} P_\theta(y|x) \quad (3.12)$$

模型预测得到的最有可能的两个类别，其预测概率越接近，说明模型越难以判断它们的类别，因此更加值得被标记。对于多分类的数据而言，仅考虑两个类别会忽略大量信息。从利用样本的所有分类结果以及对应的概率计算样本价值的角度考虑，entropy sampling 通过引入信息熵的方法，对样本的所有分类结果计算信息熵，显然信息熵较大的样本能给分类模型带来更多的改变（信息量），因此应当优先选择信息熵大的样本进行标注，如式 3.13 所示：

$$x_E^* = \arg \max_{x \in U} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \quad (3.13)$$

其中， $y_i$  覆盖所有可能的标签，这样可把握样本的整体概率信息，得到更准确的信息量。

### 3.2.2 基于多样性的采样方法

基于不确定的采样方法只考虑单个样本信息量问题，而忽视了选择样本的信息冗余问题，因此采样策略还可以从样本的多样性来考虑。若一个未标记样本与已标记样本中的样本过于接近，那么说明它与其接近的那些已标记样本具有很多相似信息，没有标记价值。因此，基于多样性的采样方法是指优先考虑那些与已标记样本集中所有样本最不相似的未标记样本，将其加入到已标记数据集中会使得该集合中样本的分布尽可能分散。常用的相似度标准有欧几里得距离、皮尔森相关系数，余弦距离等，针对关系抽取任务，从句义出发，对词向量求平均得到句向量，两个句子之间的相似度用两个句向量的余弦距离来衡量。

$$x_D^* = \arg \min_{x_u \in U} \sum_{x_l \in L} \text{CosineDis}(x_u, x_l) \quad (3.14)$$

其中

$$\text{CosineDis}(x_u, x_l) = \frac{x_u * x_l}{|x_u||x_l|} \quad (3.15)$$

### 3.2.3 基于代表性的采样方法

基于代表性的采样方法考虑未标记数据集中整体数据分布，选出最具有代表性，能够更好的表示样本空间的样本，以提高基础模型的区分能力，最终达到提高主动学习算法效率的目的。以图 3.3 为例，图中直线代表决策边界，正方形和三角形代表已标记的两类样本，圆形代表的未标记样本。因为样本 A 位于决策边界上，所以它的不确定性最高，但实际上样本 B 会给基础模型提供更多有效信息，这是因为样本 A 在样本分布中属于孤立点，信息密度低，而样本 B 在一定程度上拥有附近未标记样本的共性。

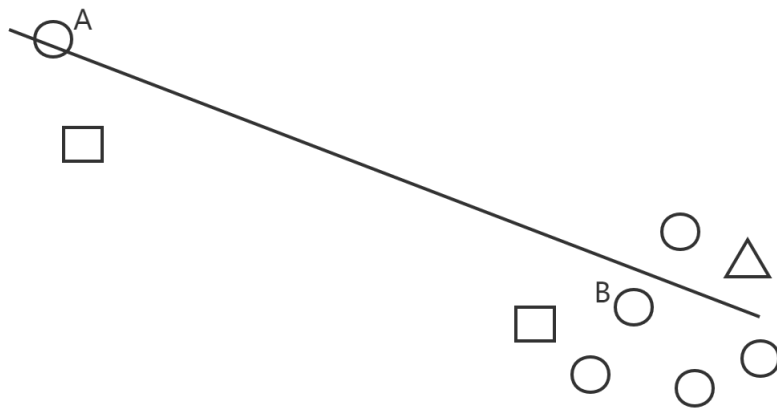


图 3.3 主动学习模型图

具体操作如下：对未标记样本集中所有样本聚类，将其划分到多个类簇中，使得相同类簇内的样本差异尽可能大，不同类簇间的样本差异尽可能小，再从中选取信息密度最大的样本，即距离类簇中心最近的样本作为本次挑选样本，其采样公式如式 3.16。

$$x_R^* = \arg \min_{x \in UC} EuclidDis(x, c) \quad (3.16)$$

$$EuclidDis(x, c) = \sqrt{\sum_{k=1}^m (x_k - c_k)^2} \quad (3.17)$$

其中  $UC$  为未标记样本簇， $x_k$  代表样本  $x$  经过基础模型特征提取之后的第  $k$  个特征， $c_k$  代表该类簇中心的第  $k$  个特征。在大多数情况下，采用高效率的 K-means 算法进行聚类，用均一性、完整性来衡量聚类效果。

## 3.3 集成采样策略

上一节提到的方法都是采用单一的挑选标准去挑选样本，然而单一的准则很难保证在所有基础模型和数据集上奏效，因而，如何有效融合多个标准来挑选样本成为我们的研究



重点。考虑到的多标准组合方式有两种。其一是通过给予不同的采样标准不同的权重对每一个样本进行评分，总得分最高的那些样本即为目标样本。其二是逐层使用不同的采样标准，上一层的筛选结果作为下一层的候选样本，不断细分候选样本，最终留下来样本是最能够兼顾多种标准的样本。

### 3.3.1 基于多标准的赋权采样策略

由于多样性和代表性的采样方法都是用来避免选择具有冗余信息的样本，从而使得已标记样本集尽可能符合全部样本空间，所以将这两部分以相等的权重结合，考虑到基于不确定性的采样方法是从样本的信息量角度出发，在以往各类任务中都被证明其具有强大的有效性，所以赋予其高权重。为有效平衡样本间的数据分布有效分散、样本内的高信息量，最终赋予基于不确定性、多样性、代表性的采样方法以 2: 1: 1 的权重对未标记样本按各自标准加权评分，总得分越高的样本，越值得被标记，算法描述如Algorithm 1。

---

#### Algorithm 1 多标准赋值采样策略

---

输入:

已标注样本集  $L$   
未标注样本集  $U$   
采样引擎  $SE$   
学习引擎  $LE$

过程:

```

1:  $Train(LE, L)$ 
2: repeat
3:    $L_f = Select(L)$  by 式 3.13
4:    $L_d = Select(L)$  by 式 3.14、式 3.16
5:    $WeightedScore(L_f, L_d)$ 
6:   for  $k = 1 : m$  do
7:      $x^* = \arg \max_{x \in U} Score(x)$ 
8:      $Label(x^*)$ 
9:      $L \leftarrow L \cup \{x^*\}$ 
10:     $U \leftarrow U \setminus \{x^*\}$ 
11:   end for
12:    $Train(LE, L)$ 
13:    $Test(LE)$ 
14: until 精度达标或超出标记代价
15: return  $LE$ 

```

---

基于多标准的赋权采样策略通过赋予不同的权重有效结合不确定性、多样性和代表性三种采样方法，并采用评分机制衡量，最终未标记样本池中分数最高的样本即为我们需要标记的样本。

### 3.3.2 基于多标准的逐层采样策略

在上一节基于代表性的采样方法的介绍中，我们选择距离类簇中心最近的样本作为该类簇的代表性样本，但实际上聚类中心的样本不一定能很好地代表该类簇整体的样本情况，受基于多样性采样方法启发，在本方法中，我们认为，与类簇内的其他样本都具有较高相似性的样本才是最能代表该类簇的整体样本。

考虑到基于不确定采样偏向于挑选最靠近决策边界的样本，这些样本能够使模型收敛加快，且不确定采样往往能够在不同任务上都具有良好的性能提升。因此，通过计算所有候选样本的不确定性，挑选不确定性最大的一些样本作为初筛样本池，紧接着，对初筛样本池聚类，在各个类簇内计算样本句向量相似度的熵值，选择熵值最大的样本为该类簇代表样本进行标注。Algorithm 2描述如下：

---

#### Algorithm 2 多标准逐层采样策略

---

输入：

已标注样本集  $L$   
 未标注样本集  $U$   
 采样引擎  $SE$   
 学习引擎  $LE$

过程：

```

1:  $Train(LE, L)$ 
2: repeat
3:    $L_f = Select(L)$  by 式 3.8、式 3.10、式 3.13
4:    $C_{all} = Cluster(L_f)$ 
5:   for  $C \in C_{all}$  do
6:      $x_c^* = \arg \max_{x_{cur} \in C} [\sum_{x \in C, x \neq x_{cur}} CosineDis(x_{cur}, x)]$ 
7:      $Label(x_c^*)$ 
8:      $L \leftarrow L \cup \{x_c^*\}$ 
9:      $U \leftarrow U \setminus \{x_c^*\}$ 
10:  end for
11:   $Train(LE, L)$ 
12:   $Test(LE)$ 
13: until 精度达标或超出标记代价
14: return  $LE$ 
    
```

---

通过多层遴选的方式尽可能地衡量每种采样标准的同时，由于每一层的输入都是经过上一层“过滤”后得到的样本集合，所以该算法的计算开销相较于基本采样策略不会提升太多。

### 3.4 本章小结

本章首先从 BERT 开始介绍 R-BERT，接着通过对现有主动学习算法的仔细调研，提出了两种融合多标准的采样策略，将多种基本采样策略有效结合，以求更佳的性能提升和场景适用，并给出算法具体描述，将在下一章实验部分验证算法的有效性、健壮性等。

## 第四章 实验及分析

本次实验有三个目的，一是验证主动学习能够在关系抽取任务中有效减少样本标记代价；二是验证提出的基于多标准采样算法与基础主动学习算法相比，能够取得较优或相当的性能。三是要验证提出的主动学习算法是模型和任务无关的或者说是可以跨任务和跨模型的。

首先，在每个数据集上划分为训练集和测试集，再将训练集划分为包含少量已标记数据的集合以及包含大量未标记数据的集合，其次，实现了多个基本采样策略，并跟随机采样在相同条件下进行对比实验，比较模型在不同策略选择出来的训练集上的性能表现。在每一次迭代过程中，算法依照不同的采样策略从未标记集合中挑选样本进行标记查询，将查询得到的标记数据从未标记集合中移除并加入到已标记集合中，接着基于已标记集合重新训练学习模型，并在测试集上对模型进行评估。

### 4.1 实验环境与数据

本次实验环境设置如下：

- 操作系统：Windows 10
- 编程语言：Python 3.6
- 深度学习框架：PyTorch 1.2
- 机器硬件：TITAN X, 12G 显存

实验数据采用百度大脑开源的 Information Extraction 数据集<sup>[7]</sup>，该数据集包括 50 个关系类别将其经过数据预处理，包括对收集到的文本数据集进行非法字符替换，删除重复样本，剩余 20 多万个样本，样本示例如表 4.1 所示，然后从中随机采样出 5 个样本大小为 12000 的数据集，其中包含 4000 个初始已标记样本，6000 个未标记样本的训练集，2000 个已标记测试样本的测试集。主动学习训练阶段中每轮迭代从所有未标记样本中取出 200 个样本获取标记，更新所在集合，迭代总计 20 次。

表 4.1 数据样式 (source: Baidu Information Extraction Dataset)

<i>sentence</i>	<i>relation</i>	<i>head</i>	<i>head_offset</i>	<i>tail</i>	<i>tail_offset</i>
《离开》是由张宇谱曲，演唱	歌手	离开	1	张宇	6
李治爱的是萧淑妃生的雍王李素节	母亲	萧淑妃	12	李素节	20
蔡长风（1910—2001），江西省吉水县人	出生地	蔡长风	4	江西	19

## 4.2 实验评价指标

在关系抽取任务中主要采用 F1 分数作为评估准则，F1 分数通过对准确率和召回率的加权求和，能够有效衡量模型性能。

$$F_1 = 2 \times \frac{precision * recall}{precision + recall} \quad (4.1)$$

主动学习旨在尽可能减少语料标记所带来的标记代价，常见的主动学习评价指标是指基础模型达到最佳性能时，该采样策略与随机采样相比，标注样本数减少的百分比，公式如下：

$$P = \frac{N_{random} - N_{active}}{N_{random}} \times 100\% \quad (4.2)$$

式 4.2 中  $N_{random}$  为随机采样达到最佳性能所需要的最少样本， $N_{active}$  为主动学习策略达到最佳性能所需要的最少样本。但在实际应用中，模型需要在所有样本全部标记的基础上训练，才能得到最佳性能，但如果需要查询标记所有样本，则违反了主动学习的主旨，即没有达到减少标记代价的目的。因此，通过假设查询次数固定且偏少，将在主动学习策略下模型所能达到的性能相较于随机采样的提升作为衡量此算法的标准，是一种更好的衡量主动学习有效性的指标，如下式所示：

$$P = \frac{P_{active} - P_{random}}{P_{random}} \times 100\% \quad (4.3)$$

式 4.3 中  $P_{active}$  代表经过少数几次主动学习策略查询之后的模型性能， $P_{random}$  代表经过少数几次随机采样查询之后的模型性能。通过观察少量采样之后的模型性能能够快速得到不同采样方法的优劣之分。

## 4.3 实验结果与分析

为了验证本章提出的基于主动学习的各类策略的实际应用效果，实验环节一共设置三组对比实验，每组实验均设置了对照实验，比较不同主动学习算法在查询过程中的精度曲线。

实验组一：针对 5 种不同的基本采样算法，随机采样作为基准对照，设置对比实验，实验结果如图 4.1。

其中，(a) - (e) 表示多种基础采样策略在 5 个不同数据集上的学习曲线，其中 entropy、lc、margin 分别代表基于不确定的采样方法中的 entropy sampling 算法、least confident 算法、margin sampling 算法、relations 代表基于代表性的采样方法、similarity 代表基于多样

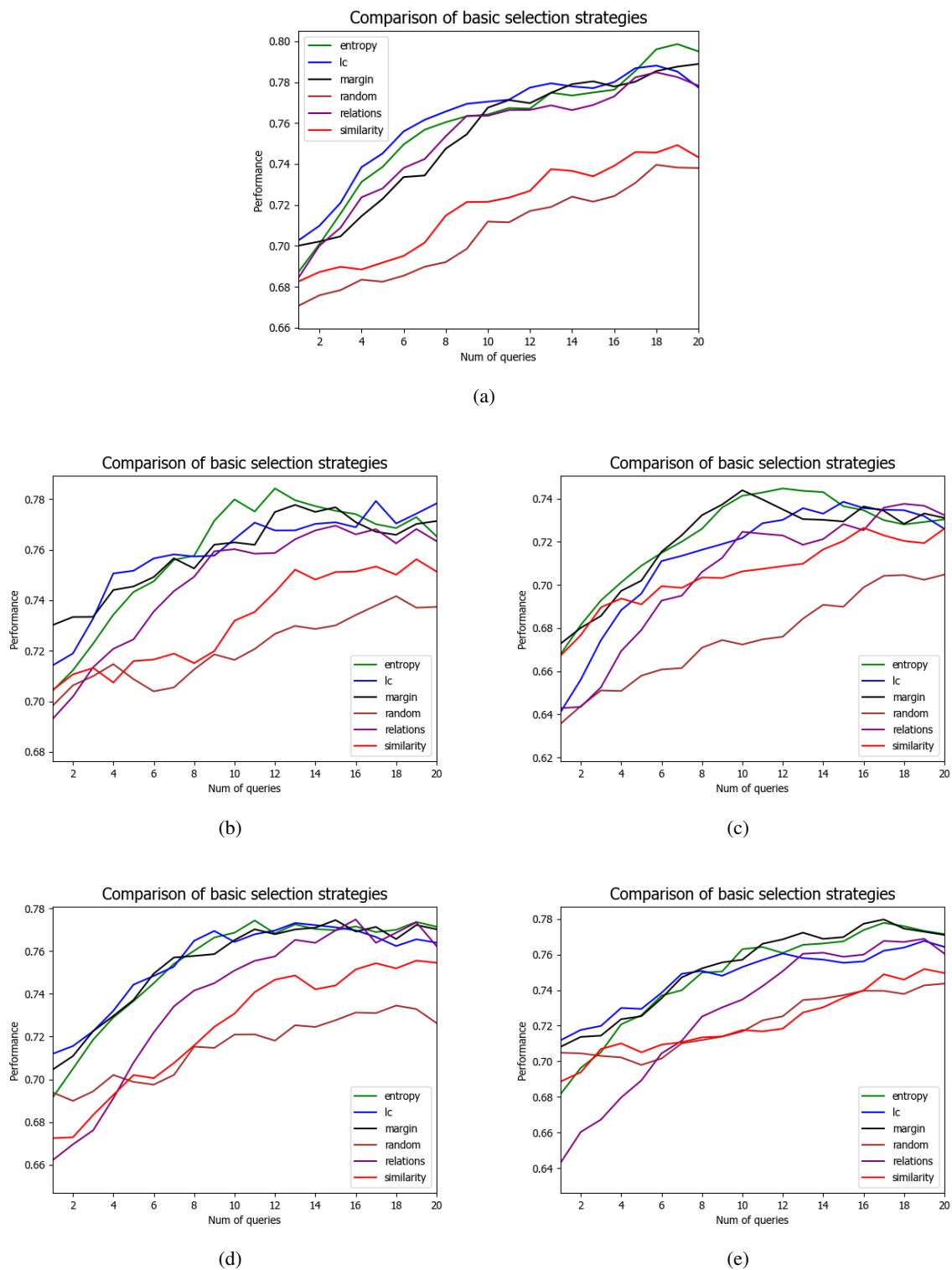


图 4.1 基本采样策略在不同语料上的对比结果

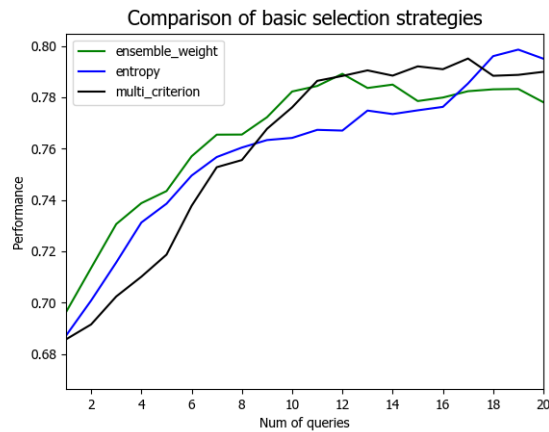
性的采样方法、random 代表随机采样。横坐标表示查询次数，纵坐标表示基于 R-BERT 基础模型的性能提升情况，在相同的查询次数时，性能表现更优的说明其采样策略更佳。

在数据集 (a), (b), (c) 上，所有的主动学习算法都明显优于随机采样，其中基于不确定的三种采样方法最为突出，验证了基于不确定的采样方法不仅在其他分类任务中奏

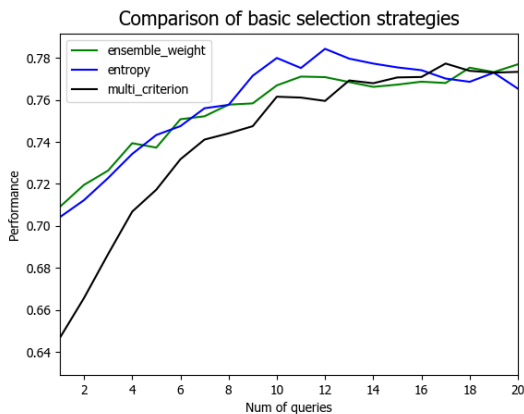
效，而且也适用于关系抽取任务。其中，**entropy sampling** 算法从信息熵的角度利用样本的所有可能分类信息，效果最佳。在数据集（d），（e）上，虽然基于多样性和代表性的采样方法在前几轮查询迭代中比随机采样差，但在之后的表现中强势上升，最后的效果依然优于随机采样，验证了通过避免样本间的信息冗余可以省去不必要的标记代价。综合多个数据集得到的分析结果，可以验证主动学习策略在关系抽取任务是卓有成效的，在关系抽取任务中，不确定的提升性能最佳，其中又以综合考虑所有类别概率分布的基于不确定的 **entropy sampling** 算法表现最优。

实验组二：针对 2 种不同的集成策略，性能最佳的基本采样策略作为基准对照，设置对比实验，实验结果如图 4.2。

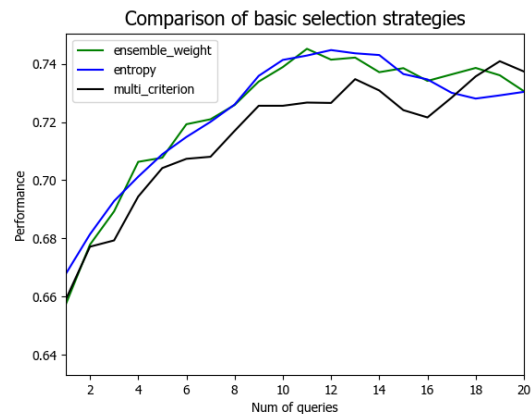
其中，**ensemble\_weight** 代表基于多标准的赋权采样策略，**multi\_criterion** 代表基于多标准的逐层采样策略，**entropy** 代表基本采样方法中最佳的基于不确定的采样方法中的 **entropy sampling** 方法。横纵坐标代表意义与实验组一相同，通过比较查询次数相同时，哪种策略性能表现更佳。



(a)



(b)



(c)

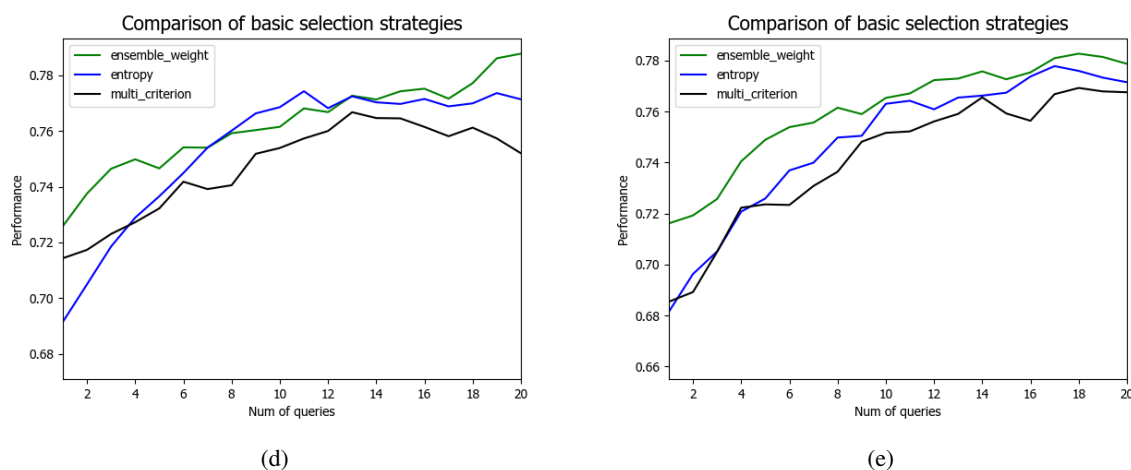


图 4.2 多标准采样策略在不同语料上的对比结果

在数据集 (a), (d), (e) 中, 基于多标准的赋权采样策略表现出优于 entropy sampling 的性能, 在数据集 (b), (c) 中, 与 entropy sampling 性能接近。通过 2: 1: 1 赋权综合衡量多种采样标准的策略在考虑了样本间数据分布有效分散的同时, 还兼顾了样本内的高信息量, 能够在多个数据集上都表现出不差于单基本采样策略的性能, 由此可以证明其是有效的。然而, 基于多标准的逐层采样策略在大部分数据集下的性能表现略差于 entropy sampling, 由此说明不是所有的多标准策略都可以奏效, 有些集成策略甚至没有单标准的效果好。

实验组三: 在同一数据集上对比三种不同的基础模型在单最佳基本采样算法、基于多标准的赋权采样策略和随机采样上的性能表现, 以验证模型无关性。对比结果如表 4.2。

表 4.2 不同模型实验结果

基础模型	随机采样	最佳基本采样策略	多标准采样策略
CNN	0.4312	+15.38%	+ 18.51%
Att-BLSTM	0.5574	+8.45%	+ 11.76%
R-BERT	0.7305	+6.31%	+ 7.68%

尽管在不同基础模型上, 数据增幅不同, 但可以看出, 不论在什么样的基础模型下, 主动学习策略都能奏效, 并且多标准采样策略都能够取得比最佳采样策略更优的性能表现, 验证了本文提出的主动学习采样策略具备健壮性。

通过三组实验, 得到多种主动学习算法在不同数据集不同基础模型上的效果, 验证了其有效性、健壮性, 达到了本文预期的设计目标, 并且验证了提出的多标准采样策略优于最佳基本采样策略, 进一步提升基础模型性能表现, 减少标记代价, 但同时, 也存在模型



变得更复杂，计算开销增加，有可能仅得到次优解的问题。

#### 4.4 本章小结

我们首先展示了我们设计的基于不确定性、多样性、代表性的主动学习算法能够比一般的随机采样在一个数据集上有更优的表现，接着我们验证两种集成策略的有效性，最后，在多个数据集上做的迁移实验表明该算法具有可迁移性和鲁棒性。

## 第五章 总结与展望

作为构建知识图谱中的重要一环，实体关系抽取的主要任务是抽取出隐藏在句子中的实体关系。目前流行的几种有监督的关系抽取模型中，基于预训练的 BERT 语言模型的关系抽取效果最佳。本次毕业设计面向中文文本，使用基于 R-BERT 的关系抽取模型为基础模型，研究基于主动学习的关系抽取策略，利用少量标注语料达到大规模、高精度标注语料的效果。

本文设计了一种基于主动学习的关系抽取方法，通过大量对比实验，验证其在不同的数据集，不同的学习模型上都具有良好的样本抽取性能。具体地，我们针对关系抽取这个任务，设计了基于不确定性、多样性、代表性三种不同的抽取策略。从百度大脑 Information extraction 中文数据集中随机采样出 5 个规模为 10000 的训练集，其中包括 4000 个已标记数据，每次根据不同的主动学习策略采样 200 个交给人类专家标注。接着，通过比较查询迭代 20 次之后的模型性能得到——基于不确定性的抽取策略最为有效，相比于随机采样，提升了 6.31% 的 F1 分数。

为了进一步提升抽取性能，通过多种集成方法综合利用原有的三种策略，最终通过实验表明，为基于不确定性、多样性、代表性三种采样策略赋予 2: 1: 1 的权重可以在原有最佳模型的基础上，提升 7.68% 的性能。

本文设计了适用于关系抽取任务下的主动学习策略，实验证明该策略具有较好的抽取效果，但仍然存在不足之处，未来的工作会从以下几个方面进行改进：

1. 本文的实验中，数据集来源较为单一，考虑到目前开源的中文关系抽取数据集较少的原因，可以通过采用第三方或者自实现爬虫系统来廉价获取大量的未标记数据。
2. 对主动学习而言，初始已标记样本的选择直接影响模型的训练时间和抽取性能。本次毕业设计中初始样本是随机采样而来，没有注重样本质量，容易导致样本分布不合理，进而影响主动学习算法的效率和最终的分类准确率，因此，对初始样本的选择方法可以考虑改进。
3. 构建基于主动学习的实体关系抽取模型中，还可以设计更优的选择策略。例如考虑文本相较于其他数据格式的特殊性，针对语义设计主动学习策略；或者采用元学习的方式，通过学习训练一个采样模型，来挑选有效样本。

## 参考文献

- [1] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]. Proceedings of Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 2004. 22–es.
- [2] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods[C]. Proceedings of Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005. 419–426.
- [3] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of machine learning research, 2003, 3(Feb):1083–1106.
- [4] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]. Proceedings of Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004. 107–110.
- [5] Wu J, Sheng V S, Zhang J, et al. Multi-label active learning for image classification[C]. Proceedings of 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014. 5227–5231.
- [6] Hakkani-Tür D, Riccardi G, Gorin A. Active learning for automatic speech recognition[C]. Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 4. IEEE, 2002. IV–3904.
- [7] Rink B, Harabagiu S. Utd: Classifying semantic relations by combining lexical and semantic resources[C]. Proceedings of Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010. 256–259.
- [8] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction[C]. Proceedings of Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005. 724–731.
- [9] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. 2014..
- [10] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks[C]. Proceedings of Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015. 39–48.
- [11] Santos C N d, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[J]. arXiv preprint arXiv:1504.06580, 2015..
- [12] Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classification[C]. Proceedings of Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016. 756–765.
- [13] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), 2016. 207–212.
- [14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Proceedings of Advances in neural information processing systems, 2017. 5998–6008.
- [15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018..
- [16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Proceedings of Advances in neural information processing systems, 2017. 5998–6008.
- [17] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]. Proceedings of Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 2361–2364.

- [18] Huo L Z, Tang P. A batch-mode active learning algorithm using region-partitioning diversity for SVM classifier[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(4):1036–1046.
- [19] Settles B. Active learning literature survey[R]. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.
- [20] Sun A, Grishman R. Active learning for relation type extension with local and global data views[C]. Proceedings of Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. 1105–1112.
- [21] Angeli G, Tibshirani J, Wu J, et al. Combining distant and partial supervision for relation extraction[C]. Proceedings of Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. 1556–1567.
- [22] Huang S J, Jin R, Zhou Z H. Active learning by querying informative and representative examples[C]. Proceedings of Advances in neural information processing systems, 2010. 892–900.
- [23] Hsu W N, Lin H T. Active learning by learning[C]. Proceedings of Twenty-Ninth AAAI conference on artificial intelligence, 2015.
- [24] Lourentzou I, Gruhl D, Welch S. Exploring the Efficiency of Batch Active Learning for Human-in-the-Loop Relation Extraction[C]. Proceedings of Companion Proceedings of the The Web Conference 2018, 2018. 1131–1138.
- [25] 朱红斌, 蔡郁. 基于主动学习支持向量机的文本分类 [D]. 2009.
- [26] 刘康, 钱旭, 王自强. 主动学习算法综述 [D]. 2012.
- [27] Al Rahhal M M, Bazi Y, AlHichri H, et al. Deep learning approach for active classification of electrocardiogram signals[J]. Information Sciences, 2016, 345:340–354.
- [28] 姚明海, 陈志浩. 基于深度主动学习的磁片表面缺陷检测 [J]. 计算机测量与控制, 2018, 26(9):29–33.
- [29] 吴健, 盛胜利, 赵朋朋, et al. 最小差异采样的主动学习图像分类方法 [J]. 通信学报, 2014, 35(1):107–114.
- [30] Angluin D. Queries and concept learning[J]. Machine learning, 1988, 2(4):319–342.
- [31] Baum E B, Lang K. Query learning can work poorly when a human oracle is used[C]. Proceedings of International joint conference on neural networks, volume 8, 1992. 8.
- [32] Dasgupta S, Kalai A T, Monteleoni C. Analysis of perceptron-based active learning[J]. Journal of Machine Learning Research, 2009, 10(Feb):281–299.
- [33] Moskovitch R, Nissim N, Stopel D, et al. Improving the detection of unknown computer worms activity using active learning[C]. Proceedings of Annual Conference on Artificial Intelligence. Springer, 2007. 489–493.
- [34] Thompson C A, Califf M E, Mooney R J. Active learning for natural language parsing and information extraction[C]. Proceedings of ICML. Citeseer, 1999. 406–414.
- [35] Lewis D D, Catlett J. Heterogeneous uncertainty sampling for supervised learning[M]. . Proceedings of Machine learning proceedings 1994. Elsevier, 1994: 148–156.
- [36] Yang J, et al. Automatically labeling video data using multi-class active learning[C]. Proceedings of Proceedings Ninth IEEE international conference on computer vision. IEEE, 2003. 516–523.
- [37] Hauptmann A G, Lin W H, Yan R, et al. Extreme video retrieval: joint maximization of human and computer performance[C]. Proceedings of Proceedings of the 14th ACM international conference on Multimedia, 2006. 385–394.
- [38] McCallumzy A K, Nigamy K. Employing EM and pool-based active learning for text classification[C]. Proceedings of Proc. International Conference on Machine Learning (ICML). Citeseer, 1998. 359–367.
- [39] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal

of machine learning research, 2001, 2(Nov):45–66.

- [40] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks[C]. Proceedings of Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008. 1070–1079.
- [41] Baidu. Information Extraction Dataset[R]. Technical report, 2020. <http://ai.baidu.com/broad/download>.

## 致 谢

大二的时候，在一个科创中，有幸很早接触了当时死灰复燃的深度学习，惊讶于深度学习强大的表征能力它在图像、自然语言处理领域取得的巨大突破让我着迷于它背后的机理，于是我试着要走进这个领域。这不进不知道，进了才发现在深度学习的背后还有机器学习这座大山，山里面有许多奇珍异宝：主动学习、反绎学习、增量学习等等等等。一股兴奋感油然而生，于是按图索骥，慢慢开始领略机器学习的各式风采。

在大四上的时候，我跟随黄圣君教授开始了解主动学习，当时我做的课题是探索主动学习在反绎学习中的应用，不过由于反绎学习是近年来刚提出的概念，参考资料十分有限，其中不少启发式算法让实验很难做，加上刚刚接触主动学习，不得其要领，大半个学期白白流失，最后愣是没出成果，不得已忍痛割爱，照顾未来研究方向，转向关系抽取。

在这之前，关于机器学习的知识，我更多是在图像领域做尝试，对于自然语言处理这块，一窍不通，好在这两个领域有不少相同之处，慢慢的也能上手，在这个过程中，感谢张永飞老师和杨山师兄十分耐心地解答我那些让人窘迫的问题，这让我有勇气有底气去写毕业设计的开题报告。

但是，很快就遇到各种问题。中文信息抽取数据集不好收集、疫情影响导致计算资源受限、对主动学习的理解不够清晰，这些问题一度让我的实验进展十分受挫。好在黄圣君教授不厌其烦地对我谆谆教导，把控毕设正确的大方向；谢明昆师兄语重心长地告诉我当前领域的最新进展，为我指点迷津；宁鲲鹏师兄不厌其烦地听我抱怨实验进度，为我出奇划策；孙峰同学毫无怨言地帮我 `review code`，指出具体的代码漏洞。这些都是善良的人啊，我除了感谢还是感谢，非要走心的话，就是祝福好人一生平安啊！

在完成这次毕业设计的过程中，阅读了不少关系抽取文献，主动学习文献，以及相关领域的最新顶会论文。在我看来，学术研究最迷人之处在于从中领略古今学者的智慧，那些神奇的妙想，从那一个个聪明的脑袋里迸发出来，形成一股科技文明之光，推动人物质文明，精神文明的发展。

四年时光就这样从指缝间流走，悄无声息。好在青山不改，绿水长流，让我们后会有期！