基于 BERT 的中文多任务模型

摘　要

随着互联网时代的来临，在我们日常浏览的网站中存在大量的文本数据，但目前对于中文而言，在处理过程中存在着词汇边界不清的问题（例如“我爱中国”），且汉语句子没有严格的主谓宾关系等语法规则限制，使得对中文文本进行有效的处理成为一个难题。本研究旨在构建多任务学习框架，把中文分词、文本分类和 NER 这三个任务有机结合，利用它们之间的相关性和互补性来达到整体性能最优的目的。

研究以 BERT 预训练模型构建共享编码层，捕获文本全局上下文语义；结合 BiLSTM 提取序列局部时序特征，利用多头注意力机制动态融合多粒度上下文依赖，增强模型对长距离语义关系的建模能力。针对数据稀疏性问题，提出融合同义词替换与义原替换的增强策略。在损失函数设计上，采用 Focal Loss 针对不同任务调整类别权重，缓解数据分布不均衡问题；引入不确定性加权机制自适应平衡多任务损失贡献，避免任务间冲突；结合线性学习率预热策略优化训练稳定性，加速模型收敛。实验表明，该框架在联合训练中显著提升任务间协同效应：分词任务通过共享实体边界信息优化了歧义切分，分类任务借助上下文特征增强了对细粒度语义的判别，NER 任务则受益于分词与分类的联合推理，三者相互促进，验证了多任务架构在资源利用与性能提升上的双重优势。

未来研究可进一步探索基于生成对抗网络的数据增强策略，并扩展模型在低资源语言与垂直领域的迁移能力，推动中文文本处理技术向更高效、更普适的方向发展。

关键词：自然语言处理；中文分词；文本分类；命名实体识别；多任务学习；BERT；BiLSTM

Abstract

With the advent of the Internet era, there is a large amount of data in the websites we browse, but for the Chinese language, there is the problem of unclear vocabulary boundaries in the process (e.g., “I love China”), and Chinese sentences do not have strict subject-verb-object relationships and other grammatical rules, making the effective processing of Chinese a difficult problem. processing of Chinese becomes a difficult problem. In this study, we aim to construct a MTL framework, which combines the three tasks of Chinese word segmentation, text categorization and NER, and utilizes the correlation and complementarity between them to achieve the overall optimal performance.

The study constructs a shared coding layer with BERT pre-training model to capture the global contextual semantics of the text; combines BiLSTM to extract the local temporal features of the sequence, and utilizes the multi-head attention mechanism to dynamically fuse the multi-granular contextual dependencies to enhance the model's ability of modeling long-distance semantic relations. Aiming at the problem of data sparsity, the enhancement strategy of fusing synonym replacement and sense-origin replacement is proposed. In terms of loss function design, Focal Loss is used to adjust the category weights for different tasks to alleviate the problem of data distribution imbalance; the uncertainty weighting mechanism is introduced to adaptively balance the loss contribution of multi-tasks to avoid inter-task conflicts; and the linear learning rate preheating strategy is combined to optimize the stability of training and accelerate the convergence of the model. Experiments show that the framework significantly improves the inter-task synergy effect in joint training: the disambiguation task optimizes the disambiguation cut by sharing entity boundary information, the classification task enhances the discrimination of fine-grained semantics with the help of contextual features, and the NER task benefits from the joint inference of disambiguation and classification, and all three are mutually reinforcing, which verifies the dual advantages of the multi-task architecture in the utilization of resources and the improvement of performance.

Future research can further explore the data enhancement strategy based on generative adversarial networks and extend the migration ability of the model in low-resource languages and vertical domains, to promote the development of Chinese text processing technology in the direction of more efficient and pervasive.

Keywords: natural language processing, Chinese word segmentation, text classification, named entity recognition, multi-task learning, BERT, BiLSTM

目　录

第1章 绪论

1.1 研究目的与意义

伴随着信息技术以及互联网的日新月异，中文数据呈爆炸式增长，广泛应用于新闻自动标注、情感分析、自动文本识别等典型应用领域。中文数据加工作为自然语言处理（Natural Language Processing, NLP）的有机组成部分，其核心任务包括分词、文本分类及 NER。由于中文无词界、无形态、无任何语法约束、歧义现象严重，中文数据加工任务难度大，尤其是分词和命名实体识别（Named Entity Recognition, NER）极难解决。

本文提出的中文文本处理模型是基于多任务学习（Multi-task Learning, MTL）框架的多任务中文文本处理模型，将中文分词、文本分类和 NER 这三个任务用 MTL 的方式组合成一个模型，来解决中文文本处理模型中文信息处理的性能问题，尤其是在数据缺乏场景下；本文对 MTL 模型进行尝试，既能够提高模型性能，又避免传统模型多任务单独训练导致多个任务共用一个模型学习资源而造成的模型浪费情况，让多个任务协同学习发挥出多任务优势。

本研究的意义在于：在双向编码器表征法（Bidirectional Encoder Representations from Transformers, BERT）等预训练语言模型语义学习的背景下，采用多任务集成方案尝试性地解决中文处理任务之间的耦合问题，通过实证应用验证方法的有效性和稳定性，为后续更复杂的中文文本处理应用实现提供方法指导和技术参考。

1.2 国内外研究现状

中文文本处理任务包括中文分词、文本分类、NER。这三个任务是中文信息处理系统的基础性任务，是进行情感分析、机器翻译、信息检索等任务的基础。

1. 中文分词：分词实则是把一句话依据一定规则重新排列成词序列的这样一个过程，而中文和英文的单词区别那便是英文单词之间是以空格作为自然分界符的，而中文只有字、句和段是凭借明显的分界符直接进行划分的，在词之间是不存在形式上的分界符的，因此中文分词比其他语言的分词任务更具挑战性。

2. 文本分类：文本分类意为将蕴含信息的一篇文本，映射至已知类别中的一个或者几个，如新闻分类、情感分析。中文文本存在着诸如语义模糊、长文本之类的问题，故而需对文本中的关键信息予以提取并展开分类。

3. 命名实体识别（NER）：NER 指的是对文本中具备特定意义的实体加以识别，例如人名、地名、组织名、专用名词等等。简而言之，也就是针对自然语言文本中的实体边界与实体类型进行分类的操作。

1.2.1 中文分词

在 NLP 技术方面，中文明显落后于西方语言，大多数用于西方语言的方法都不能直接应用于中文，因为中文必须有一个分词的步骤，中文分词从字典匹配到统计到深度学习，中文的分词还有很长一段的路要走，中文分词的技术有待很长时间才能让中文分词技术更好地服务于更多的产品。

1. 基于词典匹配：基于某种策略对被分析的中文语句与一个足够大的机器词典中的词条进行匹配，如果词典存在该词，则匹配成功。优点就是速度快，但在处理模棱两可或无法识别的词语时很吃力。

2. 基于统计建模：即根据大量的已分好词的文本，通过统计机器学习技术，从大量的预分割文本中学习到词语边界，从而对未知文本进行分割。统计方法的出现，此后隐马尔可夫模型（HMM）[1]中文分词方法被广泛应用。HMM 是对词语的统计建模，发现词语的边界，但不能很好的处理长依赖关系和上下文信息。

3. 基于深度学习：基于深度学习的中文分词近年来的性能较好，基于长短时记忆网络（LSTM）的 RNN 可以识别其中的时间信息，提高分词的准确率。基于深度学习的中文分词主要采用 LSTM-CRF 模型，提高了分词的准确率和鲁棒性，BiLSTM-CRF[2-3]模型可以较好的利用上下文语境信息标记词的边界。

1.2.2 文本分类

文本分类是 NLP 中的难题，早期方法是通过人工分类，后来发现到底以什么特征来判断文本分到那个类别，人类自己都不好回答，其中太多“可意会，不可言传”的东西。人类判断是靠经验和感觉。于是很自然地就想到了人类也是通过大量的同类的文档观察自己总结经验为以后分类的依据，就像机器一样。

1. 基于人工规则：它采用专家的规则来分类，运用知识工程构建专家系统，好处是显而易见的，它比较直观地解决了问题，但弊端也是很明显的，例如分类的好坏要看规则的好坏，再如制定规则的人都是专家，成本太高，而知识工程的最大弱点是

费时费力。

2. 基于深度学习：随着深度学习的提出，卷积神经网络（CNN）和长短期记忆网络（LSTM）等模型开始成为文本分类的主流，需要人工预先分类好的文档作为训练材料，模型可以自己自动地从这些文本中学习到一些可以对这些文本进行分类的规则，无需人工输入即可自主学习这些规则，在长文本、复杂语义等方面有更好的效果。特别是预训练模型的出现（主要是 BERT[4-6]），文本分类任务的性能有了很大程度的提升，BERT 预训练得到深层次的语义信息后，通过微调适应下游任务，在多个任务上有优良的表现。

1.2.3 命名实体识别（NER）

NER 对理解并处理文本中的关键信息十分重要，是 NLP 应用得以有效执行的关键步骤。在英语中，命名实体的形式特征（即第一个字母大写）使实体边界识别相对而言较为简单，任务侧重点主要放在识别实体类别上。相比之下，汉语命名实体的识别任务要复杂得多，且汉语相对于实体类别标注子任务而言实体边界更为难以识别。

1. 基于规则与词典：语言学家们通常会利用数据集的特点，人为设计特定的规则模板或者词典。这些规则可能包括关键字、位置词、方向词、中心词、指示词、统计信息和标点符号等。词典由特征词和外来词典组成，外来词典即现有的常识词典。规则和词典建立后，一般是根据匹配的方法对文本进行处理，进而完成命名实体的识别。这种方式虽在一些特殊的领域很有用，但是需要人工的参与，而且对复杂的或者模糊的识别能力差。

2. 基于机器学习：随着统计学习方法的逐步发展，以机器学习为基础的 NER 方法逐渐成为主流，使用人工标注的语料库进行学习，可以用于不同的领域。只需使用新领域语料进行极小的调整，就能适应新任务。典型的机器学习方法有隐马尔可夫模型（HMM）、最大熵模型（ME）、支持向量机（SVM）和条件随机场（CRF）等。虽然这些方法在不同领域都表现出色，但它们依赖于手工设计的特征，对于复杂的上下文可能不如深度学习方法有效。

3. 基于深度学习：近几年，深度学习的发展在 NLP 方面取得了较大的发展，命名实体识别开始走向基于深度神经网络（DNN）的研究。深度学习与普通的机器学习不同之处在于不需要人为提取特征，不需要专家知识，深度学习只需从海量标注数

据中学习出好的表示，并且对于许多 NER 任务都远超手工提取特征方法。深度学习成功的原因在于其强大的表示能力，以及数据中存在的多种复杂数据模态。其中 Bi-LSTM + CRF[2-3]模型是命名实体识别的成功典范。Bi-LSTM 学习文本数据中的前因和语义，CRF 修正优化文本序列标签，通过学习深度学习模型自动提取文本的有效特征大大提高了 NER 的识别率。进一步，随着预训练模型（如 BERT）的兴起，使得 NER 在各个领域上的应用取得了较大的进步，其中 BERT 模型已经成为 NER 以及其它 NLP 任务的基线模型，BERT 模型可以通过预训练海量语料，从而对文本中的实体以及文本语义有更好的理解，提高 NER 的识别率[4,7-8]。

1.3 研究内容与论文结构

第 1 章主要介绍了本设计的研究背景、目的及意义，并对中文文本处理的基本任务及 MTL 在中文文本处理的运用作了简要的概述。

第 2 章主要介绍了本研究的总体框架，包括数据标注和数据增强技术、数据预处理流程、多任务学习框架、参数共享机制、性能评估指标的具体实现。

第 3 章对本文使用的模型结构和算法进行了详细的研究，包括 BiLSTM、多头注意力机制、损失函数设计和动态学习率设置算法。

第 4 章介绍了中文分词、文本分类、NER 的实现细节与算法设计，并分析了几种算法的组合效果，最后通过实验对比验证模型的优势与不足。

第 5 章对本文的总结与展望部分，并提出了未来中文文本处理技术的改进方向。

第2章 多任务总体框架设计

本章提出多任务学习框架的整体架构，系统阐述了数据标注和数据增强技术、数据预处理流程、多任务学习框架、参数共享机制、性能评估指标的具体实现。总体框架如图 2-1 所示。

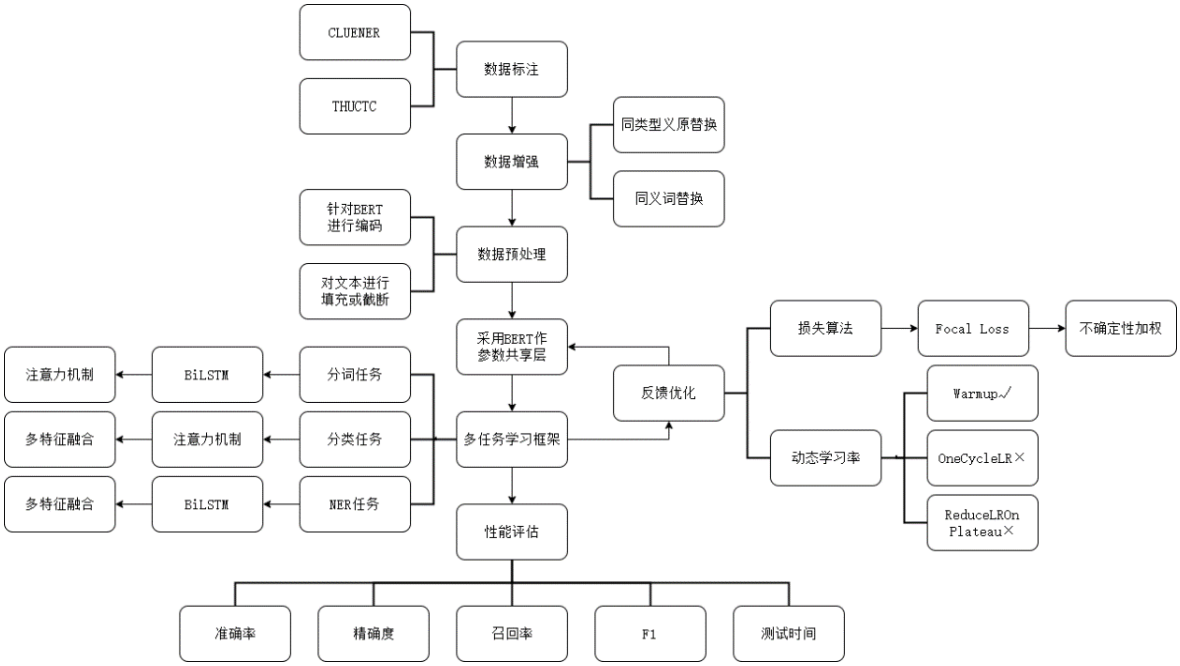

图 2-1 总框架图

2.1 数据标注

数据标注就是对原始多媒体数据（图片、文本、音视频资源）进行加工，准确标注后的样本数据集不仅可用于训练阶段，指导模型学习如何在纷繁复杂的多样性数据中进行有价值的知识模式提取，还可用于验证阶段，对模型进行评估。

toutiao-text-classfication-dataset（简称今日头条）数据集，来源于今日头条客户端，共有 382688 条数据，15 个新闻类，包括：民生类、文化类、科技类、军事类等，具有新闻类别多、文本数据丰富等特点。每个类别下的数据量分布不均，如娱乐和体育的数据量较大，而股票类别的数据量较少。

今日头条数据集每行为一条数据，以_!_分割的字段，从前往后分别是新闻编号、分类类别编号、分类名称、新闻语句和新闻中的关键词。原始数据集格式如下：“6554095647887196686_!_109_!_news_tech_!_京东回应平台出售假茅台：运输路途被调包，已报警_!_超市官方,京东,飞天茅台,茅台酒,供货商”。不难发现，尽管今日头条数据集在分类方面很完善，不仅有多个分类类别还有新闻关键词，但没有分词及

NER 的标注。在尝试标注了一部分数据集后，由于时间原因无法在规定时间结束前完成数据的标注任务，只能放弃再去寻找其他数据集。

CLUENER 是一个专为中文 NER 任务设计的大规模基准测试数据集合，共包含 12091 条数据，内含 10 种主要实体标签，如人物名、地名及机构名称等，如表 2-1 所示。该数据集相较于今日头条数据集而言，其已标注 NER 且分类标注可以在其母数据集上寻得，同时数据量较少足以完成标注任务，故而采用该数据集作为源文本及 NER 的标注，分词和分类任务均在此基础上进行修改。

表 2-1 NER 的标签类别定义和标注规则

标签类别定义	标注规则
地址（address）	**省**市**区**街**号，**路，**街道，**村等（如单独出现也标记）。地址应尽可能详细
书名（book）	包括小说、杂志、作业本、教科书、教材、地图册、烹饪书、电子书和商店出售的书籍
公司（company）	**公司，**集团，**银行（不包括中央银行和中国人民银行等政府机构），如新东方以及新华网/中国军网等平台
游戏（game）	常见的游戏包括改编自小说和电视剧的游戏，要分析具体场景判断最终是否归类为游戏
政府（government）	行政机构包括中央和地方两级。中央行政机关包括国务院及其各部门（包括各部、委员会、中国人民银行和审计署）、国务院直属机构（如海关、税务、工商、环保总局等）以及军队等
电影（movie）	电影包括一些在影院上映的纪录片和由书名改编的电影，要根据场景着重区分是电影名称还是书籍名称
姓名（name）	通常指人名，也包括宋江、武松和郭靖等小说中的人物，以及小说中的绰号（如及时雨、花和尚）和与特定人物相对应的著名人物的别称
组织机构（organization）	包括体育团队、乐队、俱乐部以及小说中虚构的帮派，如少林寺、丐帮、铁掌帮、武当、峨眉
职位（position）	包括如巡抚，知州，国师等的古代职称和如总经理，记者，总裁，艺术家，收藏家等的现代职称
景点（scene）	包括如公园、动物园、海洋馆、植物园、江河等常见的旅游景点

由于 CLUENER 是在清华大学开源的文本分类数据集 THUCTC 基础上选出部分数据进行命名实体标注后构建的数据集，THUCTC 标注的分类种类共有 14 种。因此我通过检索 THUCTC 来匹配相应的类别标签，以此来标注文本分类任务，并选取其中体育、娱乐、彩票、房产、教育、时政、游戏、社会、财经共 9 种作为实验数据集中的分类类别。

至于分词部分的标注，我决定采用 jieba 进行标注，jieba 目前在 github 上的关注数已经达到 34.1k，足以体现 jieba 在开源社区的非常受欢迎，而在实际应用中 jieba 分词器也具有良好的分词效果和较高的处理效率。

而后将分词标注结果与 NER 标注结果进行验证，确保 NER 中的实体词在分词中同为一个实体词，避免自动化标注出现的误差。

将三种标注合并后每行为一条数据，从前往后分别是源文本，分词标注，分类名称，NER 标注。最终确定了 8145 个符合实验需求的数据，其中 7244 条作为训练集，另 901 条则作为验证集与测试集。

“BIO”标注法是以“B”标注该字是词的开始，以“I”标注该字是词的内部部分，以“O”标注该字不属于任何词。这样标注便于了解文本的内容以及分词关系，有助于分词及 NER 识别。利用“BIO”标注法如表 2-2 所示。

表 2-2 数据标注样例

字符	实际分词标签	实际 NER 标签	字符	实际分词标签	实际 NER 标签
这	O	O	彩	I	I-organization
也	O	O	推	B	O
是	O	O	荐	I	O
张	B	B-name	中	O	O
路	I	I-name	第	B	O
老	B	B-position	四	I	O
师	I	I-position	次	I	O
在	O	O	命	B	O
新	B	B-company	中	I	O
浪	I	I-company	头	B	O
的	O	O	奖	I	O
足	B	B-organization	!	O	O

2.2 数据增强

数据增强是指将原始数据进行一定的改变或转换来生成新的数据，目的是通过少量的先验知识生成更多相似的数据来扩充训练集，防止模型过拟合，提升模型的泛化

能力。文本任务中一般采用同义词替换、语义等价替换等来数据增强。

2.2.1 模块介绍

同义词替换[9]指将句子里的部分词语替换成同义词，组成新的例句，在句子原有语义不变的前提下增加模型的泛化能力和稳定性。例如：“今天的天气真好”中的“天气”一词可以替换成为“气候”，“真好”一词可以替换成“非常好”，组成新例句：“今天的气候非常好”。

同类型义原词语替换通过替换具有相同义原（具有相同语义）的词语，既可保证文本的一致语义，也可提升模型对同类型实体的识别率。例如“北京”，可替换成“上海”或“广州”。

为保证数据增强的效果，以原句与替换后的句子的相关度为数据增强效果，采用计算两个句子的余弦相似度（Cosine Similarity）[10-11]，即计算两个句子的文本相似度，确保新文本与原文本语义相近，避免产生毫无关系的文本样本，确保增强的有效性、合理性。

2.2.2 方案介绍

将输入语句按分词结果分割为多个词语，对每个词语进行操作如图 2-2 所示。

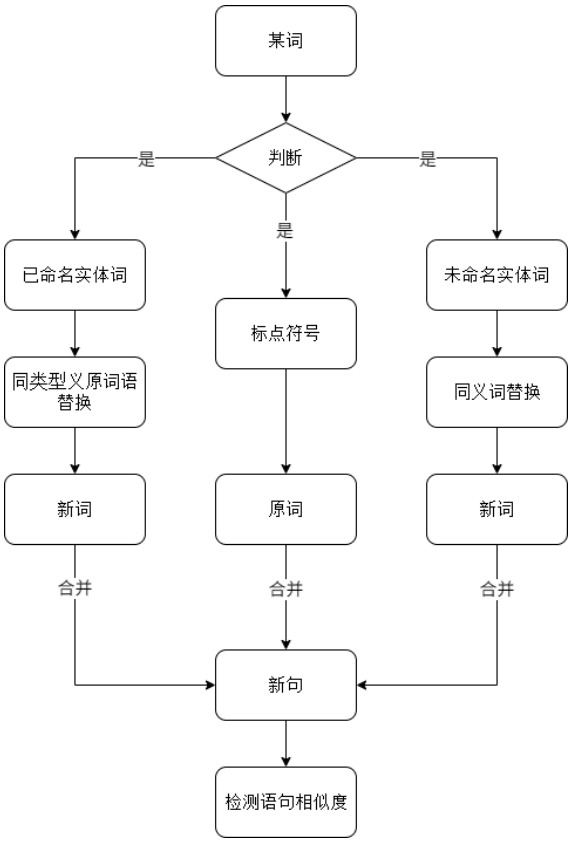

图 2-2 数据增强流程图

如果该词语是标点符号或单个字符，则直接返回词语本身，因为标点和单字通常没有同义词的需求。

如果该词是已命名实体词，根据实体类型从预先设置好的映射中获取该类型的核心义原，以计算原始词语与含有这些核心义原的所有词语的语义相似度，确保相似度在95%以上。同时检查词性一致性，并获取原始词语的所有义原，遍历之前获取的所有词语，确保新词与原始词语至少有两个义原相同。

如果该词是未命名实体词，则为其寻找多个最接近的同义词，并加以判断确保其与原词相似度在95%以上。

后将处理后的词语合并成新句子，确保新句子与原句相似度在97.5%以上，力求新生成的句子仍然保持语法正确且语义通顺。

2.3 数据预处理

数据预处理包括对原始数据进行转换，以提高其价值、减少噪声并提取有意义的信息等。数据预处理的目标是将原始数据转化为有用数据，以便后期对数据进行分析与建模。数据预处理对BERT模型是非常重要的，因为BERT模型对输入格式要求很高。

2.3.1 模块介绍

在具体使用过程中，针对BERT模型，采用专门定制的分词器将文章分词成一个个token id，对文章分词后进行编码，将语句转换成BERT模型能接受的形式，每一个token都被映射为词表ID，生成固定长度的长向量。对于BERT而言，所有输入样本必须具有相同的长度，为了保证输入文本变成固定长度，需要将文本进行填充（padding）或截断（truncation）。

2.3.2 方案介绍

将数据集中当前索引下的文本读入，将分词标注标签和命名实体识别标签从字符串类型转化为列表形式，使得文本、分词标注标签、命名实体识别标签长度一致。

对文本进行编码，将文本变为模型能接受的输入格式，根据偏移映射（offset mapping）将标签与 token 对齐。将无效的 token（如[CLS]、[SEP]和[PAD]）设置标签为-100，表示该token不参与损失计算。我提前获取文本的最大长度让BERT将文本都填充而不用截断。将标签填充为所有标签中的最大长度，不足则填充-100，让每

个样本的输入长度一致。

2.4 多任务学习框架

多任务学习（MTL）是一种不同于单任务学习的机器学习方法。在传统的机器学习中，重点是一次学习一个任务，而 MTL 则是一种联合学习，多个任务并行学习，让它们的结果相互影响。MTL 的设计假设所有任务（至少部分任务）都是相互关联的，利用相关任务的相关信息来加强所关注任务的学习过程。

目前 MTL 大致可分为两大类，一类涉及在不同任务间共享相同的参数，另一类侧重于挖掘不同任务间的共有数据特征。本次研究采用的是第一种方法，即在任务间共享参数，同时为每个任务设计各自的模型结构。

MTL 的优势在于解决了单任务模型存在的信息孤岛问题，共享表示可以提升多任务上的性能，比如基于 BERT 的多任务学习模型在文本分类、NER 等多个 NLP 任务上都有不错的表现[4-5]。通过对 BERT 的共享表示进行预训练，使得模型能达到多任务上的好表现，而不是在单任务上学差了。

2.5 参数共享机制

2.5.1 参数共享机制介绍

本研究采用参数共享的方式实现 MTL，现如今的参数共享可分为两种主要机制：硬参数共享和软参数共享。

在利用硬参数共享的多任务神经网络中，不同任务共享大部分底层隐藏层和参数，包括输入层、卷积层、全连接层等用于特征提取和表示学习的模块。这些共享层的任务是从输入数据中提取出通用的底层特征模式，捕获任务间的共同本质。不过，每个任务在网络顶层都有自己专用的输出层，用于特定的预测和决策。硬参数共享的示意图如图 2-3 所示。

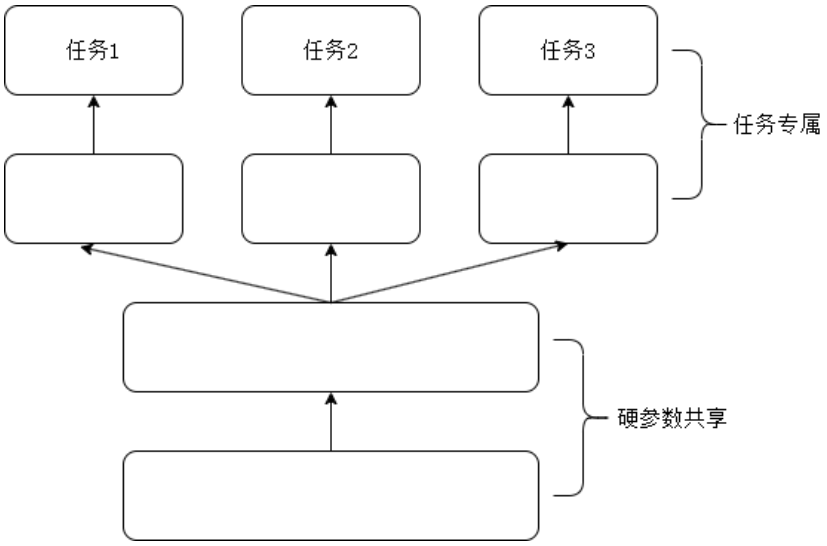

图 2-3 硬参数共享参考图

硬参数共享的主要优势在于其高效和简洁。由于大多数网络层和参数都是共享的，因此训练和部署多任务模型的计算成本仅略高于单任务模型，从而大大节省了计算资源。此外，这种共享方式也有助于防止过度拟合，提高模型的鲁棒性。

缺点是，硬参数共享可能会限制每项任务充分发挥其独特性。不同的任务可能需要捕捉不同的特征，过于严格的参数共享可能会阻碍模型拟合每个任务的能力。为解决这一局限性，软参数共享提供了一种更灵活的 MTL 架构。

在软参数共享方法中，每个任务都有自己独立的模型结构和专属参数，而不是直接共享大部分网络层。这种设计考虑到了不同任务可能需要学习不同的特征表示，允许每个任务保持一定的专属表征能力，从而提高整体性能。

不过，为了促进知识迁移和通用表示的学习，软参数共享并没有完全分离每个任务的模型参数。相反，它通过应用一定的约束来鼓励不同任务参数间的相似性，从而实现一定程度的参数共享。软参数共享的示意图如图 2-4 所示。

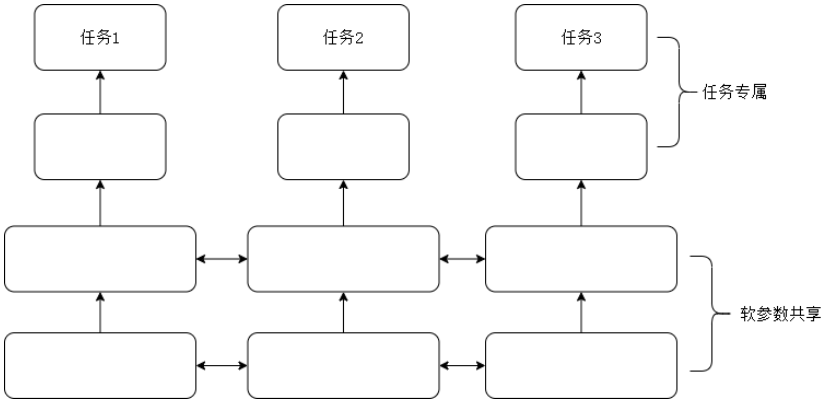

图 2-4 软参数共享参考图

通常情况下，软参数共享会在损失函数中加入一项特殊的正则化项，以评估并最小化不同任务模型参数之间的差异。这种正则化策略迫使每个任务在学习任务特定表征的同时，还要考虑其与其他任务的参数相似性，从而在一定程度上实现参数空间的共享。考虑到本研究种本研究中的分词、分类和 NER 三种任务的强关联性，最终选择了硬参数共享来实现。

2.5.2 参数共享层模型选择

BERT 是基于 Transformer 架构的预训练语言模型[4-5,12]，能够抽取文本的深层语义，对文本中的上下文关系进行建模的双向编码器，具有语言理解能力。BERT 既是一个 Word Embeddings 的生成器，也是一个 LSTM 一样的特征抽取器，如图 2-5 所示架构，将文本嵌入后的结果经 Transformer 的 Encoder (即图 2-5 中的 Trm)多层抽取特征。

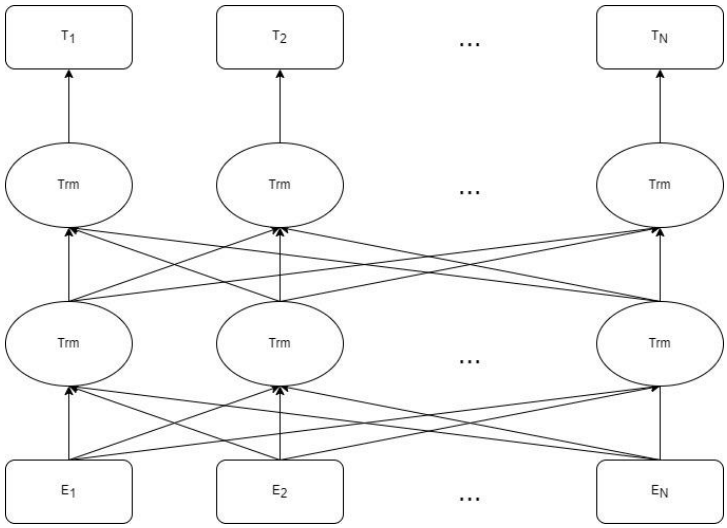

图 2-5 BERT 结构图

BERT，GPT 和 ELMo 的结构图如图 2-6 所示。

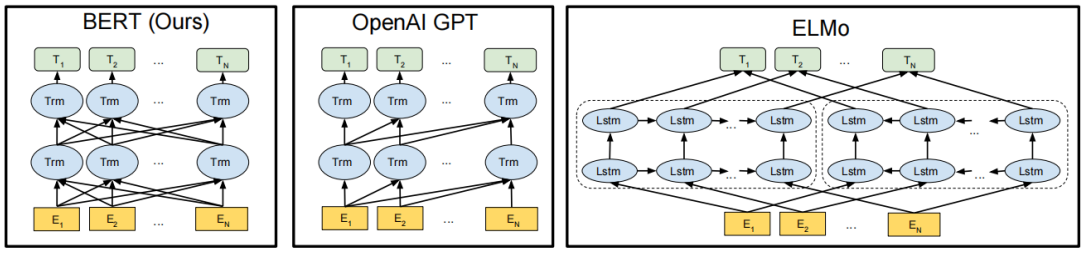

图 2-6 BERT、GPT 和 ELMo 的结构图

从特征提取器方面来看，ELMo 使用的是 LSTM，而 GPT 和 BERT 用的都是 Transformer，只不过前者是用 decoder 而后者用的是 encoder。ELMo 使用的 LSTM 提

取语义特征的能力不如 Transformer。因此在特征提取方面，GPT 和 BERT 都要更好。

从单双向来看，GPT 是单向，后两者是双向。显然，只能利用上文信息预测某个单词，不如后者利用上下文信息“完形填空”效果好。因此，我选用 bert-base-chinese 作为基础模型，在保证其专注于中文任务的前提下减少参数量。该版本一共有 12 层，隐藏层输出向量维度为 768(即 hidden size=768)。

2.6 性能评估

在评价模型中，实体的正确预测必须满足实体边界正确和实体类型正确。本文采用的评价指标分别是精确率（Precision，简写为 P）（找的对）、召回率（Recall，简写为 R）（找的全）、F1 值，如表 2-3 所示：

表 2-3 混淆矩阵

真实结果	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

其中，TP 是指预测结果为正，且真实结果也为正（即真阳性）；FP 是指预测结果为正，但真实结果为负（即假阳性）；FN 是指预测结果为负，但真实结果为正（即假阴性）；TN 是指预测结果为负，且真实结果为负（即真阴性）。F1 值为精确率（P）和召回率（R）的调和平均值，用于平衡这两个指标以达到综合最佳结果。以下是三个指标的计算公式：

$$P = \frac{TP}{TP + FP} \times 100\% \tag{2-1}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{2-2}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{2-3}$$

这三个评价指标的取值范围为 0~1，值越大性能越好，用 P、R、F1 值可以更加全面地评估模型性能。

2.7 本章小结

本章选取今日头条数据集和 CLUENER 数据集，最终选取 CLUENER 数据集，使

用 THUCTC 和 jieba 进行标注，使用 BIO 标注法，得到最终满足实验要求的数据集 8145 条；将其中 7244 条划作模型训练集，其余 901 条作为验证集、测试集，使得模型在训练集拥有足够样本量的同时有较好的泛化能力，有可靠的评价标准。数据采用同义替换、同类型义原替换增强方式，使用余弦相似度（≥97.5%）过滤结果，经过文本编码对齐填充、截断等操作后得到最终数据集，确保进入模型中的 BERT 输入统一格式，排除无关噪音。

第3章　多种模型和反馈优化算法研究

本章基于第二章提出的总体框架，对本研究使用的模型和算法进行了详细的研究，包括 BiLSTM、多头注意力机制、损失函数设计和动态学习率设置算法，旨在优化分词、分类和 NER 任务的性能。

3.1 模型算法

3.1.1 双向长短期记忆网络（BiLSTM）

LSTM（Long Short-Term Memory）是 RNN（Recurrent Neural Network）的一种，如图 3-1 所示。因为 LSTM 的设计结构使其适用于处理时序数据建模。而双向长短期记忆网络（Bidirectional Long Short-Term Memory, BiLSTM）是由前向 LSTM 与后向 LSTM 组合而成，是一种对序列数据上下文捕获的模型[2-3]，如图 3-2 所示。

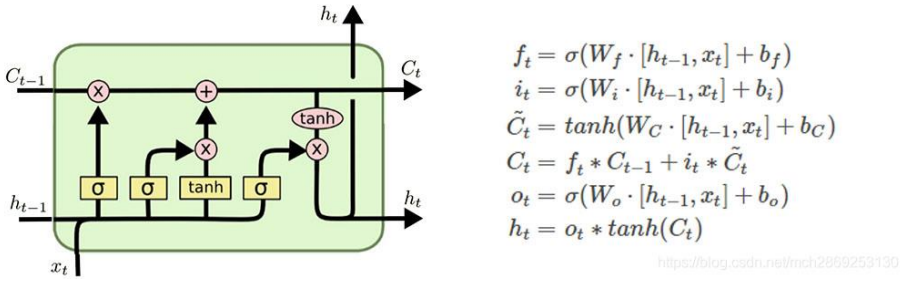

图 3-1 LSTM 结构图

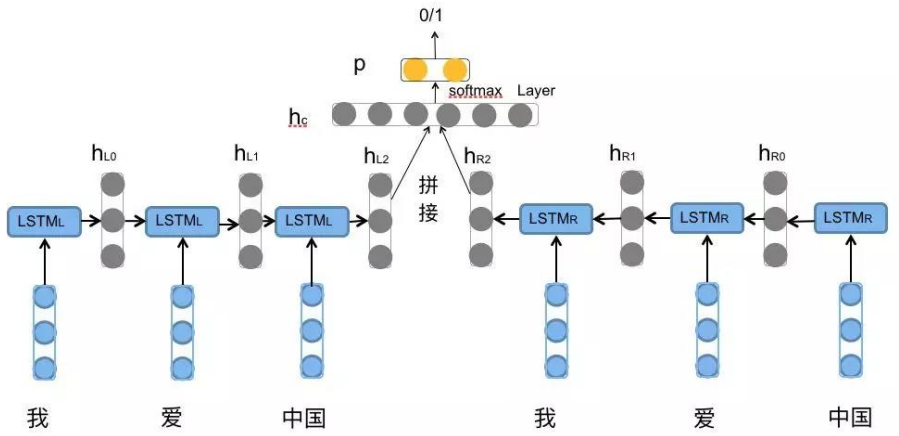

图 3-2 BiLSTM 结构图

在对句子建模时，LSTM 面临着无法从两个方向捕捉信息的限制。而 BiLSTM 则能有效捕捉双向的语义依赖，因此更适合对长文本进行上下文建模。故而在分词和 NER 任务中，我选择 BiLSTM 来处理词序列的上下文依赖性。

BiLSTM 的参数仅有隐藏层大小，因为 BiLSTM 是双向 LSTM，所以实际的输出层大小=隐藏层大小*2。对于隐藏层的大小，我们尝试了 256、384、768 之后，发现 256 特征表达能力不足（F1 下降 1.2%），768 增加了训练时间（+10%）但提升不大（F1 仅提升 0.3%），最终选择 384（即 hidden_size // 2）。

3.1.2 多头注意力机制（Muti-Head Attention）

Muti-Head Attention 是一种将输入特征拆分成多个“头”注意力机制，独立处理每个 head 得到的。Muti-Head Attention 的结构如图 3-3 所示。

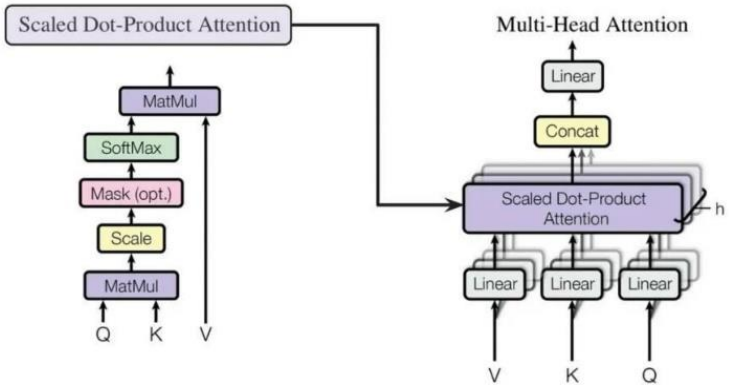

图 3-3 Muti-Head Attention 结构图

实际上，Multi-Head Attention 可以被看作是数个并行的 Self-Attention 模块，每个 head 能捕捉序列中不同的关联。这样的架构有利于强化模型学习到更长的依存关系。

Multi-Head Attention 可调的参数就只有头数。对于头数，我们分别尝试了 8 头、16 头、24 头，发现 8 头会导致特征表达能力低（F1 下降 0.7%），24 头增加了训练时间（+20%）但提升不多（F1 仅提升 0.4%），最终选择 16 头在计算效率和性能间达到平衡。

3.1.3 自注意力模型（Transformer）

Transformer 是完全基于自注意力机制（Self-Attention）处理序列数据[5,13]。这使得 Transformer 在长序列数据上的并行度更高、性能更好。Transformer 的结构如图 3-4 所示。

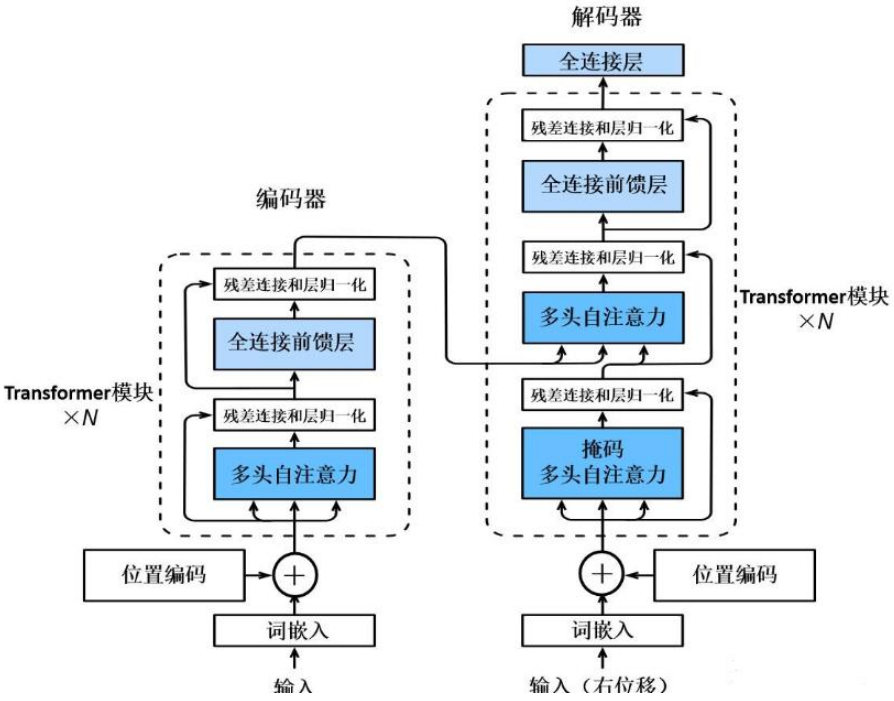

图 3-4 Transformer 结构图

Transformer 由 Encoder-Decoder 架构组成。编码器（Encoder）提取输入序列中给定序列的高阶语义，解码器（Decoder）输出序列。Transformer 的 Self-Attention 使得模型处理序列数据时，序列的任意两点间的相互联系都可以同时得到考虑，对于检测序列中的复杂依赖至关重要。

LSTM 还存在长记忆丢失问题（梯度消失），严格来说，LSTM 引入的 C 通道（见图 3-1 所示）只是缓解了梯度消失，并没有完全解决该问题，而 Transformer 引入的注意力机制彻底解决了长记忆丢失问题。注意力机制会依次遍历计算序列中任意两个词的相关度，那么即使两个词相隔很远，也能捕捉到二者的关系，从根本上解决了长时依赖难以建立的问题。但由于注意力机制的计算复杂度较高，通常为$O(n^2)$，n为序列长度，这意味着提升模型表达能力时也增加了模型的训练复杂度，所以此类模型通常都比较笨重。

Transformer 与传统的顺序输入法不同，它可以同时计算一次性输入一个序列的所有词，从而提高了并行化程度，并允许批量的对多个序列进行计算。但是，这种方法无法对序列中单词的相对位置的建模，而这对于依赖一维空间排列的语言来说是至关重要的。如果不考虑词的位置信息，就会导致打乱输入序列中词的位置，导致输出语句意义不同，很难说明两种方案孰优孰劣。

由于实验中使用的数据量较少，而 Transformer 模型参数量大（如 4 层

TransformerEncoder 参数量约 30M），在小数据场景下容易过拟合，而 BiLSTM 参数更少（单层参数量约 1.2M），结合 Dropout 后在小数据集上表现更稳定。实际测试中，Transformer 在验证集上的 F1 波动幅度（±1.2%）显著高于 BiLSTM（±0.5%）。同时由于 Transformer 的时间复杂度为$O(n^2)$，而 BiLSTM 为$O(n)$。在序列长度 64 时，BiLSTM 每 epoch 耗时约 4.7 秒，而 4 层 TransformerEncoder 每 epoch 耗时约 5.1 秒（+8.5%）。且 Transformer 的注意力矩阵显存占用较高（如 64x64 的矩阵需 32KB），在 RTX 3060（6GB 显存）上限制了 Batch Size 的进一步扩大。

3.1.4 丢失法（Dropout）

与从每个节点学习的正常神经网络不同，利用 Dropout 技术的神经网络会在训练过程中随机隐藏一些单元再进行本次训练和优化。这种随机性意味着每个批次都要训练不同的网络，从而防止了隐藏层单元间的共同适应。因此，神经元不能依赖于其他特定神经元来纠正其错误。因为“剔除”过程会确保并非所有神经元都会出现在每次迭代中。这样权值的更新有助于消除对隐藏节点间固定关系的依赖，阻止了某些特征只与其它特定特征结合在一起才有效果的情况。

对于 dropout 率，我们对比了 0.0、0.1、0.2，取 0.0 会出现过拟合现象（F1 下降 2.1%），取 0.2 会让训练收敛变慢，选择 0.1 效果最佳。

3.1.5 多特征动态融合

为了进一步增强模型的性能，本节引入了多个特征进行动态组合，这种方法可以灵活分配每个特征的重要性，以动态权重整合不同来源的信息，避免各特征简单叠加带来的信息丢失问题，使模型具备应对复杂任务的能力。

3.2 损失算法

3.2.1 焦点损失（Focal Loss）

Focal Loss 是交叉熵损失函数（Cross Entropy Loss，CE）[14-15]的改进版。它主要解决的是 Cross Entropy 的失衡问题。Focal Loss 引入了一个缩减因子使其在训练过程中区别易分辨的样本和难分辨的样本，让训练的样本偏向不易分辨的样本，让模型具备少样本学习能力，其函数表达式如式(3-1)所示。

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}log(p_t) \tag{3-1}$$

其中$\gamma \in [0,5]$，当γ为 0 时，就变为了 CE。通过 γ 可以控制简单/难区分样本数量失衡，α_t可以抑制正负样本的数量失衡，$(1-p_t)^\gamma$可以减低易分样本的损失贡献，从而增加难分样本的损失比例。

解释如下：当p_t趋向于 1，即说明该样本是易区分样本，此时调制因子$(1-p_t)^\gamma$趋向于 0，说明对损失的贡献较小，即减低了易区分样本的损失比例。对于 γ 的不同取值，得到的 loss 效果如图 3-5 所示。

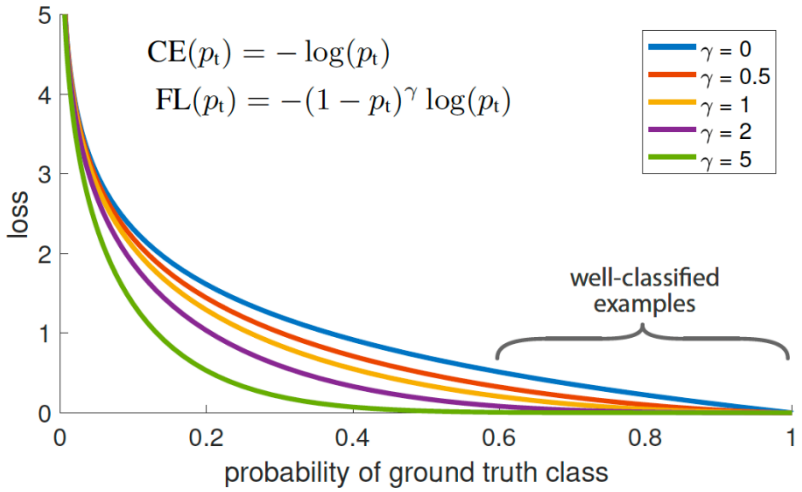

图 3-5 loss 效果图

p_t可以看出，分组越大越好分组的样本越容易分出类别的不均衡性，对于 loss 的影响就越小。对于文本分组和 NER 任务来说，Focal Loss 很好地解决了文本类别的不均衡性问题。

对比 CE Loss 和 Focal Loss 发现，CE Loss 针对低频实体（如“职位”）的召回率仅为 65%，F1 为 78.16%，而 Focal Loss（$\alpha = 0.25$，$\gamma = 3.0$），低频实体召回率提升至 69.5%，F1 达 78.35%。

Focal Loss 可调参数有α和γ。对于α，类别平衡即设定为 0.5，类别不平衡则设定为 0.25，实验中我将分词任务的α设定为 0.5，分类和 NER 任务的α设定为 0.25 以解决类别不平衡问题。对于γ比较 1.5、2.0、3.0，可发现 1.5 会略微降低对困难样本的关注（F1 下降 0.3%），但会明显增加对于简单样本的关注（F1 提升 1.5%），而 3.0 与 1.5 表现相反，会明显增加对困难样本的关注，但略微降低对简单样本的关注。最

后，对于类别平衡的分词选用 Focal Loss（$\alpha = 0.5$，$\gamma = 2.0$），对于类别不平衡的分类选用 Focal Loss（$\alpha = 0.25$，$\gamma = 1.5$），对于类别不平衡的 NER 选用 Focal Loss（$\alpha = 0.25$，$\gamma = 3.0$）。

3.2.2 不确定性加权（Uncertainty Weighting）

Uncertainty Weighting 是基于样本不确定性调整损失函数中的样本权重提出的[15-16]，通过计算每个任务的输出不确定程度并赋予不确定性高的任务较大权重，提高任务的鲁棒性，避免了任务冲突，能适应更多复杂程度不同的任务。

而固定权重策略是一种简易有效的处理方式，适用于任务差异不大且相对均衡的情况，且可人工调整固定权重来保证任务间的均衡。但固定权重的计算方法不具有自适应调整据单个任务难度或者其学习程度的能力，容易导致某些任务的表现欠佳与过拟合。

通过测试，采用 Uncertainty Weighting 后比固定权值下，分词任务的准确率上升 2.07%，分类任务的准确率上升 0.89%，NER 任务的准确率上升 1.64%。因而选择了 Uncertainty Weighting 来解决模型在不同任务间的不确定性，让某些任务不过分地依赖模型训练，在一定程度上解决了过拟合的问题。

3.3 动态学习率算法

3.3.1 学习率预热（Warmup）

Warmup 在预热期间，学习率从线性增加到优化器中的初始预设学习率，之后使其学习率从优化器中的初始学习率线性降低到 0，如图 3-6 所示。

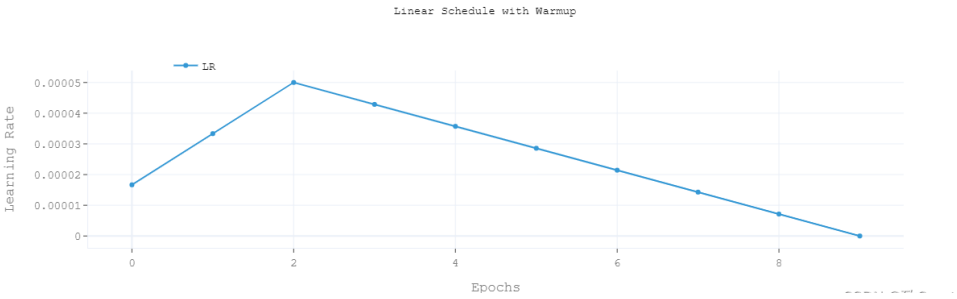

图 3-6 warmup 效果图

模型的权重在开始时是随机初始化的，此时使用较大的学习率会使模型振荡。Warmup 可以帮助我们从较低的学习率开始训练几个 epoch，让模型在应用预热后的

学习率趋于稳定，从而加快模型收敛速度且改善效果。

分别尝试了学习率和权重衰减（weight_decay），学习率分别为 1e-5、5e-5、1e-4，1e-5 收敛太慢，完全收敛额外需要 10 epochs，1e-4 导致梯度爆炸，选择 5e-5 使验证集 F1 达到最佳，权重衰减分别测试 0.01、0.05 与 0.1、0.1、0.1，0.01 表示轻过度拟合，0.1 表示欠拟合，选择 0.05 限制更新 BERT 参数，提升泛化能力、防止过拟合。

3.3.2 OneCycleLR（One-cycle learning rate schedulers）

OneCycleLR 在训练过程中，学习率在早期阶段快速升高以跨过局部最优解，然后逐渐降低以使模型更易于收敛在全局最优解处，如图 3-7 所示。

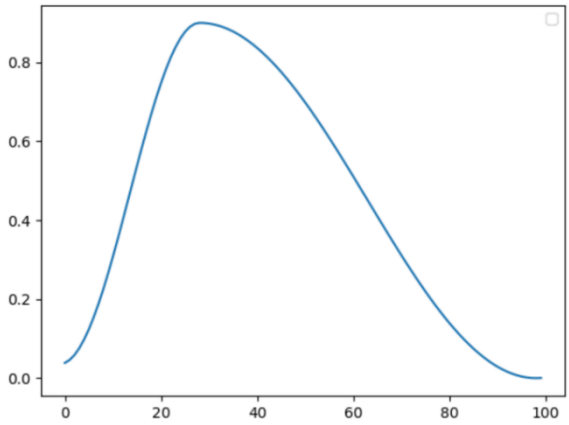

图 3-7 OneCycleLR 效果图

OneCycleLR 对比于 Warmup 同样可以加快训练速度并防止过拟合，但在数据集较少或训练时间短的情况下会存在不稳定因素，与本文的研究目标提升在有限数据情况下中文文本处理的性能不符，因此不选用 OneCycleLR。

3.3.3 ReduceLROnPlateau（Reduce Learning Rate On Plateau）

当监测模型在验证集上的性能停滞时，ReduceLROnPlateau 便会根据特定指标持续降低学习率，使模型更好地收敛，可以很好的解决模型训练后期无法收敛的问题。

经过对比实验，采用 Warmup 相比 ReduceLROnPlateau 使分词任务的准确率上升 5.06%，分类任务的准确率上升 0.44%，NER 任务的准确率上升 4.58%。发现使用 Warmup 学习率更积极，以避免因陷入局部最优解导致的提前降低学习速率，使得训练不稳定，因此放弃使用 ReduceLROnPlateau。

3.4 本章小结

本实验选择的 BERT 版本为 BERT-Base，该版本的模型一共有 12 层网络和 12 个

head，隐藏层输出向量的维度为 768。本章所提出的识别模型的训练参数设置如表 3-1 所示。

表 3-1 模型训练参数

参数	取值
Max len	64
Train epochs	20
Train batch size	64
Learn rate	5e-5
Weight decay	0.05
Dropout	0.1
hidden size	384

最初模型的分词准确率仅在 85%左右，分类准确率仅在 81%左右，NER 准确率仅在 88%左右，在不断调整模型过程中模型性能不断提升。其中 Max len 表示输入句子的最大长度，设为 64；Train epochs 表示模型训练的次数，设为 20；Train batch size 表示训练中每次处理多少条语句，设为 64；Learn rate 表示模型学习的速率，设为 5e-5；Weight decay 表示学习率的权重衰减，设为 0.05；Dropout=0.1 表示每个神经元在每次迭代中有 0.1 的概率被丢弃，该参数可防止模型过拟合；BiLSTM 网络隐藏层大小（hidden size）为 384。

第4章 多任务模型架构设计算法实现

本章将详细描述模型架构中每个任务的设计与实现，重点对比不同算法组合在分词、分类和 NER 任务中的性能，并分析选择某些算法的原因。我将对比实验结果与准确率，以进一步展示每种设计的优缺点。

4.1 实验数据集和参数设置

某次测试结果将其按顺序列出后，测试结果如表 4-1 所示。

表 4-1 数据标注样例及预测结果

字符	实际分词标签	预测分词标签	实际 NER 标签	预测 NER 标签
这	O	O	O	O
也	O	O	O	O
是	O	O	O	O
张	B	B	B-name	B-name
路	I	I	I-name	I-name
老	B	B	B-position	B-position
师	I	I	I-position	I-position
在	O	O	O	O
新	B	B	B-company	B-company
浪	I	I	I-company	I-company
的	O	O	O	O
足	B	B	**B-organization**	**O**
彩	I	I	**I-organization**	**O**
推	B	B	O	O
荐	I	I	O	O
中	O	O	O	O
第	B	B	O	O
四	I	I	O	O
次	I	I	O	O
命	B	B	O	O
中	I	I	O	O
头	B	B	O	O
奖	I	I	O	O
！	O	O	O	O

某次测试结果图 4-1 所示，其中绿色为预测正确，红色为预测错误，标签对比的格式为：某字（原标签->预测标签）。

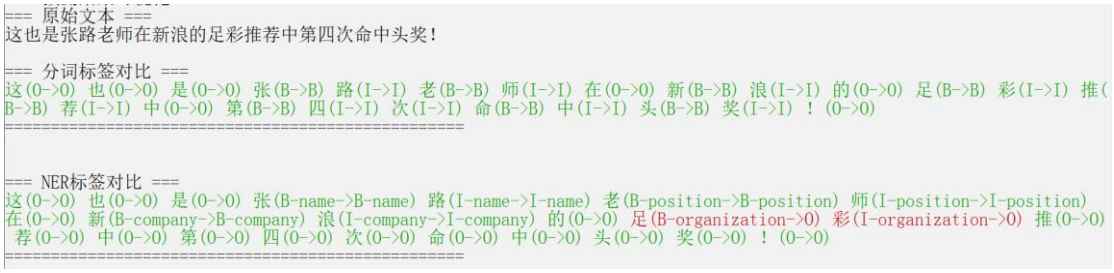

图 4-1 实验结果图

本文实验的环境如表 4-2 所示，所有代码均在 python 中实现。

表 4-2 实验环境

项目	环境配置
操作系统	Windows 10
CPU	AMD Ryzen 7 5800H @3.2GHz
GPU	NVIDIA GeForce RTX 3060（6GB）
内存	32GB
Python	3.9.21
Scikit-Learn	1.6.1
PyTorch	2.6.0+cu118

4.2 分词任务

对于分词问题，本文模型采用 BiLSTM + Muti-Head Attention[2-3,17]。BiLSTM 可以为词序列添加依赖关系，Muti-Head Attention 可以帮助网络更好地理解上下文语境。不但考虑局部连接，同时还可以使用全局注意力提升分词准确率。

此外，我们还对比了应用 Transformer + Muti-Head Attention 和 BiLSTM + Muti-Head Attention 结合多特征动态融合策略，实验结果如表 4-3 所示。

表 4-3 分词测试结果

组合	准确率	相比 3	测试时间/s	相比 3
1	91.08%	-0.39%	4.6375	-1.11%
2	91.24%	-0.21%	5.104	+8.82%
3	91.44%	0%	4.6899	0%

1. Transformer + Muti-Head Attention：Transformer 体系架构优势较多，但小数据集容易出现过拟合，中文分词不能较好的捕捉局部时序性，模型不能很好提取细粒度特征，导致整体准确率不高。

2. BiLSTM + Muti-Head Attention + 多特征动态融合：在 BiLSTM 的基础上，加入多特征动态融合层，增加了模型复杂度，降低了模型训练收敛效率，经对比验证，提升效果差。

3. BiLSTM + Muti-Head Attention：与方案 2 相比计算速度更快，方案 2 增加了模型表达能力，但是分词效果不好且增加了计算量，相较于方案 1 降低了过拟合的风险，在数据量不大的时候准确率更高。实验结果如图 4-2 和图 4-3 所示。

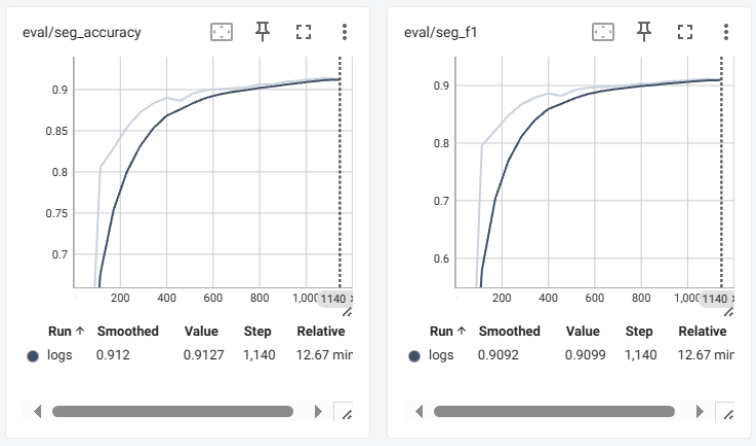

图 4-2 分词结果图 1

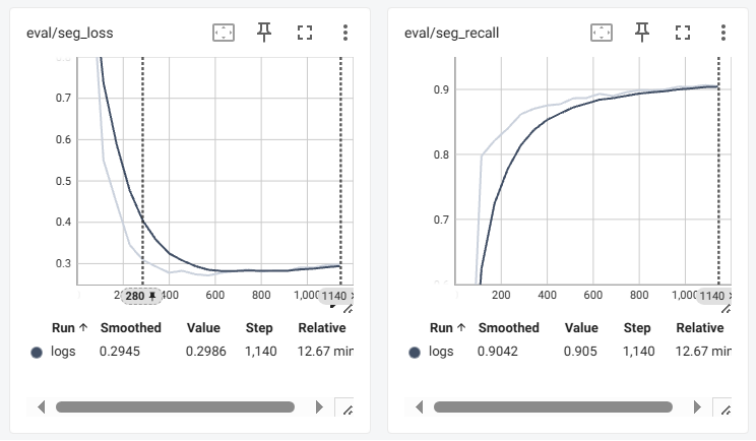

图 4-3 分词结果图 2

经由一系列对比实验验证，我们发现 BiLSTM + Muti-Head Attention 机制在执行分词任务时取得了最佳效果，实现了高精度分类。与其他设计方案相比，该组合保持了较低的计算成本。在扩增数据后再次进行分词测试，结果如表 4-4。

表 4-4 分词在数据增强前后测试结果

评估办法	增强前	增强后
P	90.51%	91.98%
R	90.37%	91.43%
F1	91.01%	91.69%

分词任务相较于增强前，分词的精度上升了 1.62%，召回度上升了 1.17%，F1 上升了 0.75%，说明数据增强后的训练样本使得模型更能学到语言的本质，能够提升词语划分的精度，一定程度上提升了模型对中文分词的认知。

4.3 分类任务

在分类任务部分，本人采用了 Muti-Head Attention + 多特征动态融合 + Linear[5,13,18]，Muti-Head Attention 学习到文本长距离信息，多特征动态融合后采用 Linear 作为分类输出，既能学习更深的文本特征，也能降低模型运算复杂度。

而在分类中，我也尝试了仅使用 Linear 层来进行分类，实验结果如表 4-5。

表 4-5 分类测试结果

组合	准确率	相比 2	测试时间/s	相比 2
1	89.78%	-0.74%	4.6899	-6.59%
2	90.45%	0%	5.021	0%

1. Linear：简单直线模型也可以达到一定的准确率，但是，简单的直线的模型不能学习到文中深层次的语义信息，太长的文书根本无法正确。

2. Muti-Head Attention + 多特征动态融合 + Linear：加入 Muti-Head Attention 模型，在全局域下建模，且多特征融合后的模型更具表征力，模型精度显著提升。实验结果如图 4-4 和图 4-5 所示。

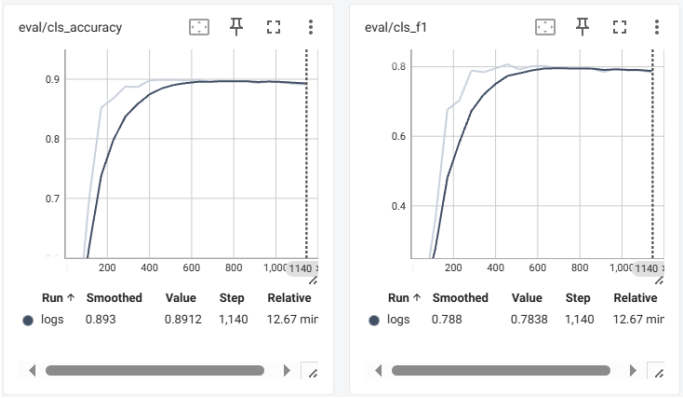

图 4-4 分类结果图 1

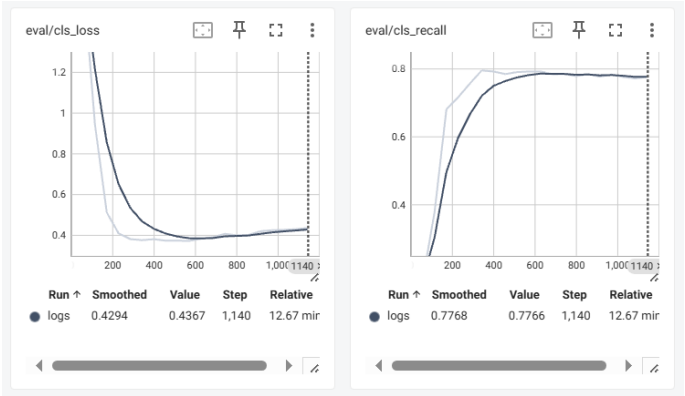

图 4-5 分类结果图 2

因此，Muti-Head Attention + 多特征动态融合 + Linear 要比仅仅使用 Linear 分类

的效果好，它不仅可以处理长文本，而且多特征的融合提高了分类任务的分类效果。在扩增数据后再次进行分类效果测试，结果如表 4-6 所示。

表 4-6 分类在数据增强前后测试结果

评估办法	增强前	增强后
P	90.45%	90.34%
R	77.65%	78.61%
F1	78.92%	80.39%

分类任务相比之前准确率降低了 0.12%，召回率提升了 1.24%，F1 提升了 1.86%，各类型如表 4-7 所示。可以发现数据增强后，模型对小样本更为稳定，但由于多数影响了准确率，反而降低了准确率。因此分类数据增强后的准确率变化不明显。

表 4-7 分类各类别在数据增强前后测试结果

分类类别	增强前 P	增强后 P	增强前 R	增强后 R	增强前 F1	增强后 F1
体育	92%	89%	84%	88%	88%	88%
娱乐	88%	91%	81%	83%	84%	85%
彩票	81%	85%	88%	85%	84%	85%
房产	90%	90%	94%	94%	92%	91%
教育	70%	80%	47%	64%	56%	64%
时政	77%	76%	80%	76%	78%	76%
游戏	95%	95%	96%	96%	95%	96%
社会	45%	58%	35%	41%	39%	46%
财经	91%	92%	95%	96%	93%	94%

4.4 NER 任务

在 NER 任务中，我采用 BiLSTM + 多特征动态融合模型[7-8,18]。BiLSTM 可以提取文本中的时间序列特征，多特征动态融合层可使模型具有更好的特征融合能力，从而提升了实体识别的性能。

除此之外，我还尝试了 BiLSTM + Muti-Head Attention + 多特征动态融合以实现 NER 任务，并将实验结果总结于表 4-8 中。

表 4-8 NER 测试结果

组合	准确率	相比 2	测试时间/s	相比 2
1	92.11%	+0.09%	5.104	+1.65%
2	92.02%	0%	5.021	0%

1. BiLSTM + Muti-Head Attention + 多特征动态融合：理论上 Muti-Head Attention 有利于模型的全局语义学习，但其在 NER 任务中不利于模型的训练，反而增加了任务的难度和训练负担。

2. BiLSTM + 多特征动态融合：它能够将文本中的命名实体进行识别，具有模型复杂度较低，准确率较高的特点。实验结果如图 4-6 和图 4-7 所示。

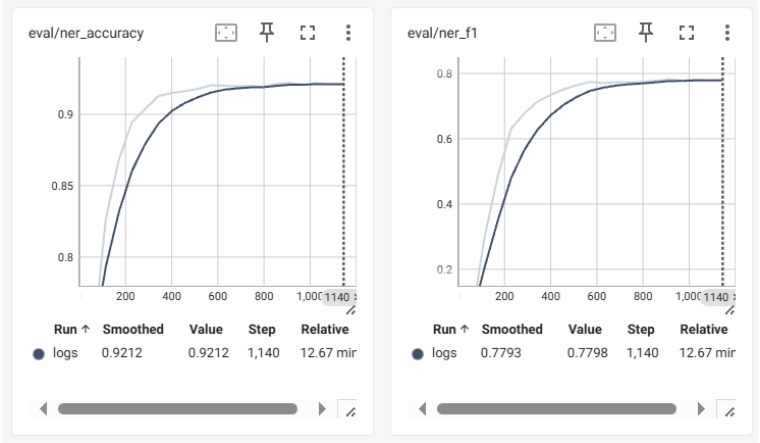

图 4-6 NER 结果图 1

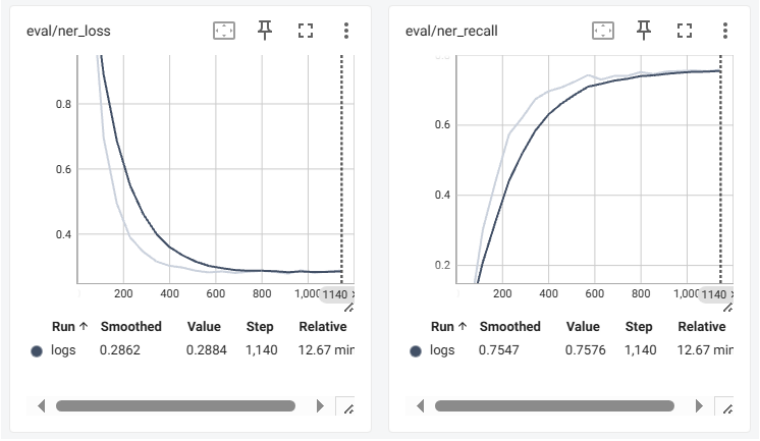

图 4-7 NER 结果图 2

通过对比，我发现 BiLSTM + 多特征动态融合在 NER 任务中表现最佳，其准确率与更复杂的设计相当，但模型复杂度较低，训练效率更高。因此，我最终选择了 BiLSTM + 多特征动态融合。在扩增数据后再次进行分类效果测试，结果如表 4-9 所示。

表 4-9 NER 在数据增强前后测试结果

评估办法	增强前	增强后
P	92.02%	92.85%
R	75.58%	76.39%
F1	78.16%	78.35%

NER 任务准确率较增强前增加 0.9%，召回率提高 1.07%，F1 提高 0.24%，说明数据增强后训练数据对模糊实体边界或上下文的改变适应能力增强，实体识别率提高。由于实体替换率不高，造成低频率的实体（如职位），与其他任务相比提升较少。

4.5 本章小结

本次研究将公开数据集 CLUENER 和 THUCTC 作为实验数据集进行研究，在经过数据增强以及模型的筛选与调参后，模型性能相较于最初模型显著提升，其中分词任务提升约 6.98%，分类任务提升约 9.34%，NER 任务提升约 4.85%，展现了多任务模型的强大潜力，验证了所提方案的有效性。数据增强在分词和 NER 任务上有很大的上升空间，但是分类任务数据增强的上升空间有限，因此在后续工作中，可以针对特定的领域进行数据增强，例如：采用特定领域的知识增强、在分类任务中引入更深层次模型架构，进一步发掘数据增强的能力。

第 5 章 总结与展望

5.1 总结

本项目的研究工作围绕中文信息处理中的中文分词、中文文本分类、命名实体识别这三个中文文本分析领域的核心内容开展了大量研究，提出多任务联合训练模型并进行实验研究。本文的主要研究工作及成果如下。

1. 针对中文语言文本处理复杂数量级的特点，文本分、类别、命名实体识别相结合的联合训练系统可以做到在文本处理系统中的不同子任务系统之间进行信息共享，使得各个组件之间协同工作。中文文本处理系统运行良好。

2. 构建了一个基于 BERT 和 BiLSTM 的网络体系结构的混合模型，注意力和 BiLSTM 的优势互补，使模型在三个基本任务的预测能力方面得到质的提升。实验证明，提出的模型在样本数量不多的前提条件下同样能够从任务的联合中体现出泛化能力强的特点。

3. 通过采用同义概念词语替换、同义词替换等方法来增强数据，使得训练集质量得到了提升，系统稳定性、鲁棒性均有所增强。特别是对分词任务，效果十分明显，而在 NER 任务上，虽有所改善，但效果甚微，也从一定程度上证明所采用方案的正确性。

4. 通过不同模型和损失函数之间的对比，通过实验确定了一种基于多特征融合和焦点损失的新的最优解，并解决了不平衡分布的数据问题，进一步优化了文本分类和名词识别的指标。

5.2 展望

尽管当前构建的 MTL 框架在中文文本场景下展现了极大的优势，但是依然存在以下几个可探索的点。

1. 数据增强策略改进：上述方法对分类问题的改进增强(0.21%)未达到预想效果，未来可从基于外部数据集库和基于领域专家知识进行数据语义增强；基于机器学习算法的智能数据增强算法：基于生成对抗网络（GAN）模型[7-8,18-19]进行数据增强，等等。

2. 跨语言迁移能力：中文较好，跨语言迁移能力较差。我们会通过迁移学习[20]

将其拓展到其他语言场景下，特别是缺少小语种的文本分析，证实该系统具备很好的跨语言迁移能力。

3. 模型透明化：虽然 BERT-BiLSTM 组合模型取得了较好的效果，但由于 BERT-BiLSTM 的组合模型是“黑箱”，模型不透明，现尝试构建一个可视化的决策路径和层次化特征解释模型，以增加模型的解释性和透明度。

4. 语境建模增强：NER 任务不能依赖数据扩充（0.9%），即更高级别的语义信息意识，引入粒度级语义语境信息，融合多粒度级信息，完成复杂语境识别。

综上，本研究为中文语境下的多任务协同文本分析提供了一种新的研究思路，后续研究中还会从语料质量、跨语言迁移性以及模型解释分析等方面开展研究，推动该技术研究更加成熟可靠。

参考文献

[1] Meng W C. Liu L C. Chen A Y. et al. A comparative study on Chinese word segmentation using statistical models[C]. IEEE International Conference on Software Engineering and Service Sciences, 2010: 482-486.

[2] 刘宇鹏, 栗冬冬. 基于BLSTM-CNN-CRF的中文命名实体识别方法[J]. 哈尔滨理工大学学报, 2020, 25(01): 115-120.

[3] Huang Z H. Xu W. & Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. 2015, ArXiv, abs/1508.01991.

[4] 陈赛. 基于 BERT 的中文命名实体识别方法研究[D]. 南京邮电大学, 2023.

[5] Fan Y X. Xie X H. Cai Y Q. Chen J. Ma X Y. Li X S. Zhang R Q. & Guo J F. Pre-training Methods in Information Retrieval[J]. 2021, ArXiv, abs/2111.13853.

[6] 习景运. 基于深度学习的中文文本分类算法研究[D]. 南昌大学, 2023.

[7] 李张岩. 基于多特征融合和数据增强的中文命名实体识别方法研究[D]. 北京化工大学, 2024.

[8] 李彦楠. 基于多特征融合和特征提取增强的中文命名实体识别[D]. 哈尔滨商业大学, 2024.

[9] Tian Y H. Song Y. Xia F. Zhang T. & Wang Y G. Improving Chinese Word Segmentation with Wordhood Memory Networks[C]. Annual Meeting of the Association for Computational Linguistics, 2020, 58: 8274-8285.

[10] Dai Z Y. Callan J. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval[J]. 2019, ArXiv, abs/1910.10687.

[11] Bai Y. Li X G. Wang G. Zhang C L. Shang L F. Xu J. Wang Z W. Wang F S. & Liu Q. SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval[J]. 2020, ArXiv, abs/2010.00768.

[12] Collobert R. Weston J. Bottou L. Karlen M. Kavukcuoglu K. & Kuksa P. Natural Language Processing (Almost) from Scratch[J]. 2011, ArXiv, abs/1103.0398.

[13] Li J. Sun A X. Han J L. & Li C L. A Survey on Deep Learning for Named Entity Recognition[C]. IEEE 39th International Conference on Data Engineering, 2023: 3817-3818.

[14] Loshchilov, I., Hutter, F. Fixing Weight Decay Regularization in Adam[J]. 2017, ArXiv, abs/1711.05101.

[15] Ren X. He W Q. Qu M. Huang L F. Ji H. & Han J W. AFET: Automatic Fine-Grained Entity

Typing by Hierarchical Partial-Label Embedding[C]. Conference on Empirical Methods in Natural Language Processing, 2016: 1369-1378.

[16] Zhang Y. Yang J. Chinese NER Using Lattice LSTM[C]. Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, 56: 1554-1564.

[17] Zhu Y Y. Wang G X. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 3384-3393.

[18] Li S. Zhao Z. Hu R F. Li W S. Liu T. & Du X Y. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]. Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, 56: 138-143.

[19] 刘爽. 基于深度学习的中文嵌套命名实体识别研究[D]. 长春工业大学, 2024.

[20] Emelyanov A. Artemova E. Multilingual Named Entity Recognition Using Pretrained Embeddings, Attention Mechanism and NCRF[C]. Workshop on Balto-Slavic Natural Language Processing, 2019, 7: 94-99.