**一、研究方案设计**

1.研究方法

针对本课题的研究内容和特性，达到研究目标和完成毕业论文，采用如下的研究方法：

（1）理论知识和技术准备：采用文献调查法，利用学校图书馆、网上相关学术数据库等资源，了解当前最新的关系抽取相关的研究进展。学习自然语言处理，深度学习以及相关技术实现，选择部分论文进行复现。

（2）设计技术方案：在理论和技术准备的基础上，结合自己的创新点，提出新的算法，设计出本课题详细的研究方案；在对设计研究方法结果分析的过程中，比较不同算法，证明其有效性和可行性。

（3）实现及测试：实现自己的算法并在公开公认的数据集上进行测试。

（4）发表一篇小论文。

（5）完成毕业论文：整理文献资料、代码和数据等，完成论文。

2、技术路线

实验步骤初步设计如下：

（1）数据采集与预处理：根据调研的论文，选择公开认可度高的数据集作为基本训练数据，在此基础上选取关系种子，在Wiki百科或者纽约时报等大型语料集上发现噪声数据集。对所得到的文本数据进行预处理，需要对文本数据进行分词、构建词表等操作，获得实体位置、词-实体相对位置等信息，对不同长度的文本数据作对齐处理。

（2）文本编码器：编码器主要包括文本特征表示、句子特征抽取和注意力层。编码器结构如图1所示。

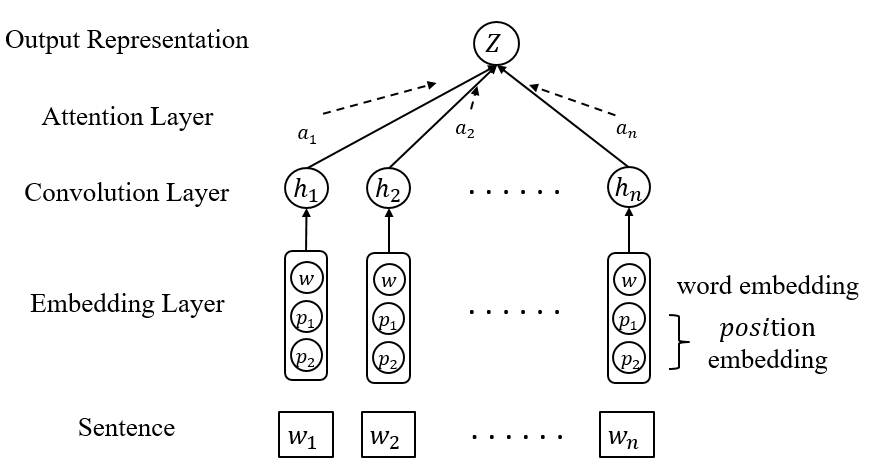


图 1文本编码器结构图

文本特征主要包括词向量特征、位置向量特征等特征。词向量方面，拟采用两种方式进行词向量表示：1）使用基于word2vec模型训练得到的词向量。2）使用基于BERT预训练得到的词向量。位置向量方基于每个词分别对两个实体的相对位置。句子特征抽取基于PCNN、RNN等模型搭建，最后引入注意力机制完成整个编码器的搭建。

（3）生成对抗网络模型构建：构建生成模型和判别模型，定义模型损失函数，设计对抗训练过程。 生成模型对从Wiki百科等语料库中发现的噪声数据进行打分，其目标函数为最大化判别模型输出概率，因此得到其损失函数计算公式如3-1所示：

|  |  |  |
| --- | --- | --- |
|  |  | （3-1） |

其中为生成模型的输出。

判别模型的损失函数由两部分构成，一部分为判别真实训练数据得到，一部分为判别自动标注的噪声数据得到，计算公式如3-2所示：

|  |  |  |
| --- | --- | --- |
|  |  | (3-2） |

生成对抗网络结构如图2所示：

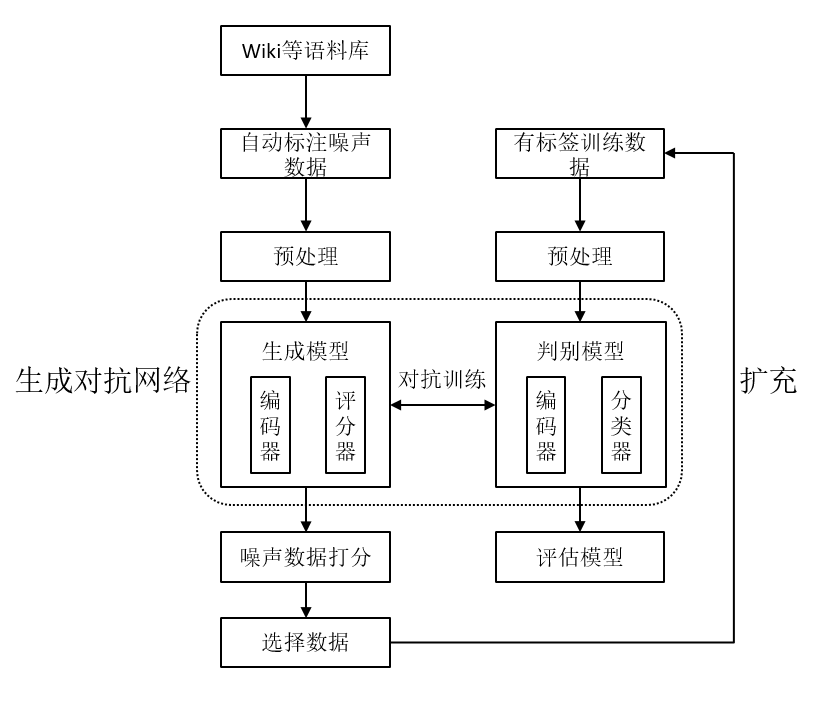


图 2生成对抗网络流程图

（4）模型效果评价：采用Precision，Recall，F1-score作为评价标准。

（5）关系抽取原型系统：在得到模型后，将其集成到关系抽取分析系统中，完成可视化结果展示。原型系统架构如图3所示。

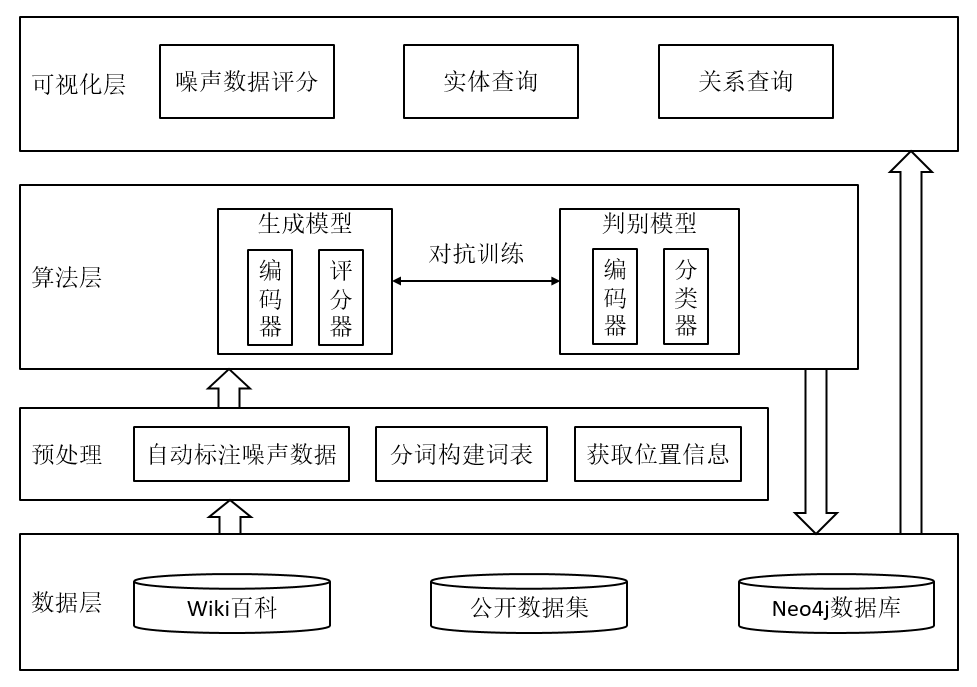


图3 关系抽取系统架构图

**二、可行性分析**

1）数据预处理。数据预处理工作可以利用NLP开源工具，或者pytorch提供的Torchtext等工具，可以快速完成文本的分词、构建词表。

2）生成对抗网络搭建。生成模型和判别模型的结构较为简单，其损失函数公式定义明确，利用Pytorch可以完成两个模型搭建。

3）模型评估。采用准确率、召回率、F1值、P-R曲线等指标进行评估，这些指标计算公式明确，且可以借助matplotlib工具进行绘制相关图表。

4）系统搭建。基于Django框架可以快速完成系统的搭建，其官方网站有详细的指导手册便于学习，可以与Neo4j数据库完成后端的连接，实现可视化效果。

论文研究自然语言领域的热点之一——实体关系抽取，主要研究基于有监督学习的实体关系抽取算法，算法主要基于生成对抗网络。开题报告检索了实体关系抽取算法的相关文献，对实体关系研究现状进行了充分调研。目前实体关系抽取算法主要分为有监督学习和远程监督学习，有监督学习需要大量人工标注的语料，而远程监督学习需要远程知识库且引入了大量噪声数据，因此论文提出的利用生成对抗网络实现训练数据集扩充并得到一个具有抗噪声的实体关系抽取模型具有很好的理论和实际意义。

关于论文课题研究内容的技术可行性：

一、生成对抗网络在图像和自然语言处理领域都有广泛应用，报告详细阐述了生成对抗网络的模型，对网络中的生成模型和判别模型的功能、架构进行了细致的描述，生成模型作为选择器是可行的且具有一定的创新性。

二、在文本特征提取方面，目前主流的方法包括word2vec、BERT等语言模型进行词向量提取，且取得了不错的效果。利用该方法进行文本特征提取是可行的。

三、注意力机制近几年在NLP领域有着广泛应用，对于过滤特征信息里的无关信息起到了关键作用，引入注意力机制的实体关系抽取模型具有可行性。

关于论文成果要求和考核指标：

选择公开，权威的数据集，测试模型的效果，以precision，recall，F1值来评价模型的效果。在数据集上具有较高的准确性和召回率。搭建实体关系抽取系统模型，将结果可视化展示。论文成果要求发表一篇EI检索以上学术论文。