本项目的核心为推荐算法的设计，通过词向量的大规模训练来实现药物的最相似匹配。模型采用Bert+BiLATM，词向量训练采用skip-gram方法。

Bert模型：

Bert（Bidirectional Encoder Representation Transformers）模型是Google发布的一个语言预训练模型。可以理解为一个通用的NLU模型，为不同的NLP任务提供支持。本项目主要采用的是bert发布的中文的一个预训练模型Bert-base-chinese，训练的数据集是基于大料的中文语料，如各种百科资料等，对于各种常用的语句都能打到比较高的理解能力。本项目实在已有的基础上进一步训练形成对古汉语和中药典籍的理解。

Bert的框架由编码器和解码器组成，编码器是由输入数据经过positional encoding进行编码，再经过multi-head attention模块进行双向计算，经过一系列的维度，映射变换得到编码器的结果。解码器的结构与编码器类似，主要功能是将编码器的结果与上一步的输出结果同时进行运算，得到最后的序列。在实际的模型中有多个类似的结构，实现语义的分析，具体的细节不再赘述[1]。

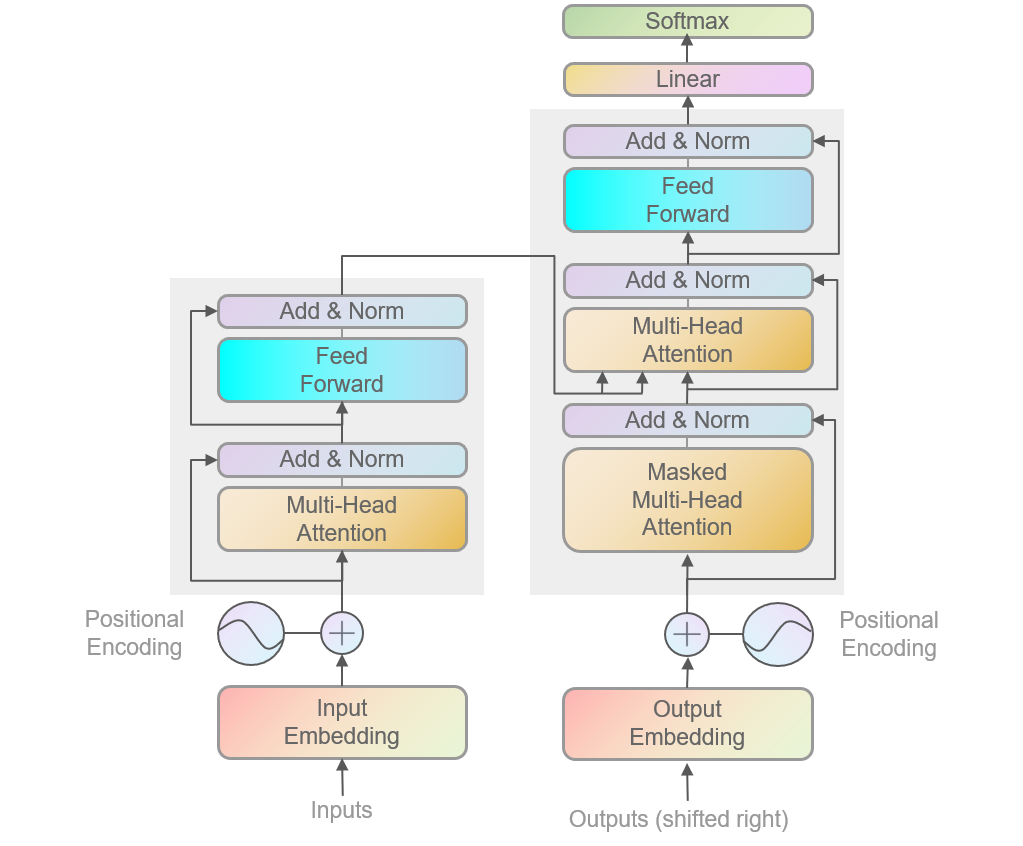


图1 Bert模型的结构[1]

Bi-LSTM模型：

双向LSTM模型是LSTM的一种改进，也是RNN模型的延伸。LSTM的主要特点是克服了神经网络中长期时间关系训练效果不好的问题，通过遗忘门、输入门、输出门的结构，在seq2seq领域大放异彩。双向LSTM解决的问题主要是LSTM只能从前向后单向计算的问题，双向循环神经网络通过增加从后往前传递信息的隐藏层来更灵活地处理这类信息。

在于给定的序列长度为t的输入（n为样本数，输入个数为d）和隐藏层激活函数。在双向循环神经网络的架构中， 设正向隐藏状态为（正向隐藏单元个数为h），反向隐藏状态为（反向隐藏单元个数为h），分别计算正向隐藏状态和反向隐藏状态：





式中，均为模型参数，通过梯度反向传播进行训练。

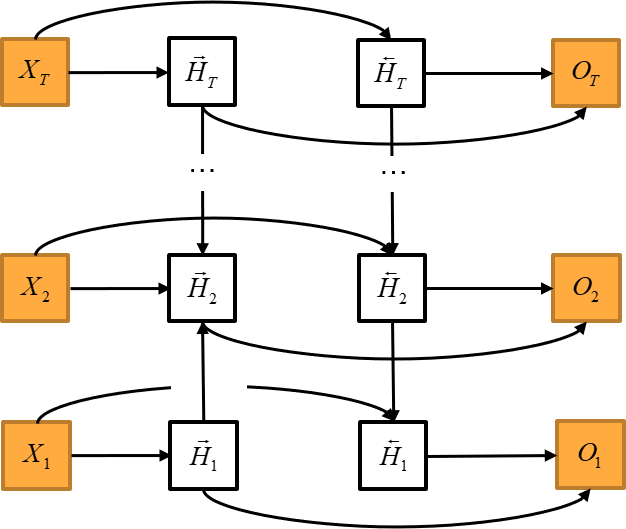


图2 双向LSTM模型示意图[2]

skip-gram方法：

skip-gram的主要用途是通过对语料进行处理，来实现词向量的精确描述的。主要方法是通过构建一个利用单个词来估计前后词的概率来作为误差进行训练。Skip-gram包含三个模块，输入、映射、输出，其构架如图所示。模型中的为输入词，在已知词的前提下预测的上下文，，，，条件概率写为：，目标函数为：



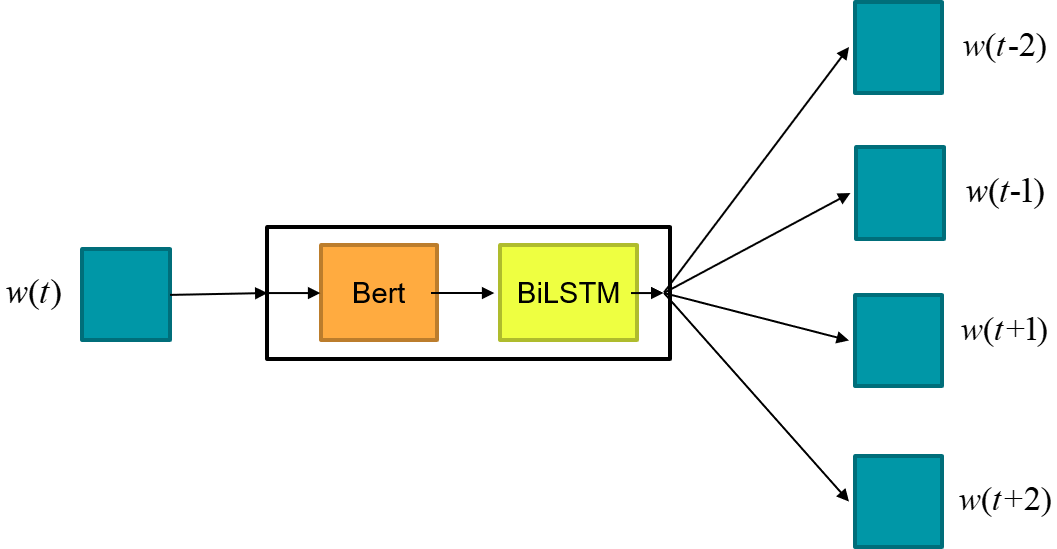


图3 Skip-Gram算法[3]

其中Bert和BiLSTM方法，都是进行双向训练，这也很符合中医的一些习惯用语，比如：脾虚胃寒跟胃寒脾虚，足手跟手足都是一个意思。另外Skip-Gram选取前后各两个词的长度作为训练目标也主要是因为许多词语都是固定的搭配，且相互之间表达的意思相近，如舒筋活络等词语。

1. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018. <https://arxiv.org/abs/1810.04805>
2. Czum J M . Dive Into Deep Learning[J].Journal of the American College of Radiology, 2020, 17( 5):637-638.<https://zh.d2l.ai/>
3. Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.<http://arxiv.org/abs/1301.3781>