Enriching Pre-trained Language Model with Dependency Syntactic Information for Chemical-Protein Interaction Extraction

Jianye Fan, Xiaofeng Liu, Shoubin Dong
South China University of technology

论文摘要

•生物医学文献中化学-蛋白质相互作用（**CPI**）的自动提取在许多生物医学应用中发挥着重要作用，例如药物发现，知识发现和生物医学知识图谱的构建。然而，复杂句子中的**CPI**是很难提取的。现有的大多数方法主要关注的是序列信息而不是句法信息，而实际上句法信息也有利于**CPI**的提取。本文提出了一种基于预训练语言模型的融合依存句法信息的方法，以提高**CPI**提取的性能。首先，该方法根据**CPI**数据的特征提取出概括性的依存句法信息。然后，采用**BERT**来生成序列信息和句法信息的上下文表示。并且使用平均池化方法来聚合上下文表示。最后，将序列信息和句法信息融合并馈入**softmax**层以获得分类结果。对原始**ChemProt**语料库的实验表明，与其他基于预训练模型的方法相比，我们的方法可以实现更好的性能。

系统模型

BERT的**Transformer**结构+信息融合层

论文简介

使用融合句子依存句法信息特征的预训练语言模型来提取生物医学文献中的化学-蛋白质相互作用(**CPI**)，在原始**ChemProt**语料库上显示了方法有更好的效果

算法原理

依存句法信息提取算法大致是：

1.对于一般情况，提取两个实体和根节点之间相连路径上的词语，组成依存句法信息

2.对于重叠实体的情况，提取该重叠实体和根节点最短路径上的词语，组成依存句法信息

3.对于两个实体位于不同句子的情况，分别取两个句子上实体与对应根节点的最短路径上的词语，组成依存句法信息

将依存句法信息和句子序列信息组合，馈入到**BERT**中，对于两部分的输出做融合操作

实验仿真

实验对比了六种方法在**CHEMPROT**数据集上的**F1**值

1）BERT模型

2）BERT + MTB。 Soares等人提出的一种方法是利用实体标记令牌来帮助基于**BERT**的关系提取。

3）R-BERT。 Wu等人开发的模型，用实体信息丰富了预训练语言模型用于关系分类。

4）BioBERT。 Lee等开发的一种预训练的生物医学语言表示模型，用于生物医学文本挖掘，他们将经过预训练的**BERT**语料库从**Wiki**更改为**Pubmed**和**PMC**，从而提高了**BERT**在生物医学任务上的性能。

5）BERT+Gaussian。 Sun等人提出的一种方法，将高斯概率分布和外部生物医学知识引入**BERT**。

6）SciBERT（**replace**）。 **Liu**等人提出的一种基于**SciBERT**的方法，用统一的单词代替目标实体。

对比实验显示我们的方法能够比之前最优的方法提升约**0.48%**

同时还做了消融实验，在我们的方法中进行了三种丢弃设置

1）丢弃基于**PubMed**的生物医学预训练模型，改用最基本的在**wiki**上预训练的**BERT**模型

2）丢弃依存句法信息部分，而是只使用句子序列信息作为输入

3）更改依存句法信息的提取方法，将只取最短路径改为取与路径上节点距离为**1**的所有节点

实验显示丢弃不同的组件均会对性能产生一定的下降影响

论文结论

本文提出了一种利用依存句法信息来增强预训练语言模型在化学-蛋白质相互作用提取任务上性能的方法。我们的方法从**CPI**数据中提取高度概括的依存句法信息，然后使用**BERT**生成序列信息表示和句法信息表示。在序列信息和语法信息融合之后，输出将被馈送到**softmax**层以获得最终的分类结果。从句子中提取的核心依存句法信息改善了模型的性能。对**ChemProt**语料库的实验表明，我们的模型优于其他基于模型的预训练方法。消融研究表明，良好的预训练模型和适当的依存句法信息有利于化学-蛋白质相互作用的提取。

将来，我们的目标是更有效地利用生物医学文本的句法信息，并使我们的方法进一步适应更多的生物医学数据集。

西安电子科技大学 XIDIAN UNIVERSITY