面向语义相似度匹配的临床术语标准化方法

1. 背景与方法调研

临床术语标准化任务是医学统计中不可或缺的一项任务。随着信息化发展，各医院累积了海量电子病历档案，如何有效利用这些数据来提高医疗水平是当前研究热点之一。临床上，由于医生个人书写习惯以及术语表达多样性等因素，关于同一种诊断、手术、药品、检查、化验、症状等往往会有成百上千种不同的写法，严重阻碍了医疗数据的检索、分析和利用。标准化（归一）要解决的问题就是为临床上各种不同说法找到对应的标准说法。。把形式多样的医疗概念映射到标准的医疗术语编码,即临床术语标准化对于疾病辅助诊疗、科研检索、疾病分组以及智能医保等研究具有重要的现实意义。因此,研究一种自动的临床术语标准化方法对于推进医疗信息化建设以及减少编码人员的工作量、提高工作效率具有重要的现实意义。

尽管临床术语的标准化编码对于医疗行业的发展非常重要,但在实际情况中患者就诊时为其分配正确的编码,不但耗时、易出错且成本高昂。医疗机构的专业编码人员会对临床医生的诊断进行二次编码,编码人员需要查阅医生对某次临床事件写的诊断描述以及电子病历中的其他信息,然后再按照编码指导以人工的方式分配合适的标准编码。这种方式总体学习成本较高,并且效率低下。因此,研究一种自动的临床术语标准化方法对于推进医疗信息化建设以及减少编码人员的工作量、提高工作效率具有重要的现实意义。

对于该类问题，本项目将该问题视作为一种文本相似度检测问题，采用相似度检测来进行检测。因此，相关可行方法也以探讨文本相似度检测问题为主。根据调研，大部分研究也是依照该类方法展开。本项目在对于文献检索，以及对于文献相关工作模块的阅读之后，概括性的整理了以下几种案例。

采用人工进行筛选：

1. Ghiasvand等在SemEva-l2014任务7上提出的基于编辑距离特征来生成候选集的方法,该方法首先通过训练数据的每一条实体及其在 UMLS 系统对应的标准术语学习到554种编辑距离模式,该模式之后被应用到测试集来增强候选答案的覆盖,在SemEval任务上取得了最佳性能;
2. Kang 等[6]在生物领域文本上提出了 5 种规则来提升疾病术语的归一化性能。

采用机器学习进行筛选：

3） 宁温馨利用分布式语义相似度计算方法计算文本语义相似度,并且基于词语和汉字构建词向量来提高性能。

4） 黄嘉俊主要基于领域知识库结合分词、实体识别和词向量表示技术进行临床术语相似度计算。

5） Larkey等通过组合三种分类器:K-近邻、相关性反馈以及贝叶斯独立分类器,对住院患者的出院记录实现临床术语自动标准化。

采用神经网络进行筛选：

6） Shi提出了一种分层神经网络模型,可以捕获到原始词与标准词之间的潜在语义信息,并且设计了注意力机制来解决原始词对应标准词数量和预测的标准词数量之间的不平衡问题。

7） 赵提出了一种基于深度学习的电子病历医学实体标准化算法,使用 Siamese网络架构和 LSTM 网络搭建模型,采用 Pairwise方法训练模型。

8） 张等构建了一种基于融合条目词嵌入和注意力机制的深度学习模型,该模型首先对含有病案条目的文本进行融合条目的词嵌入,并通过关键词注意力来丰富词级别的类别表示,然后利用词语注意力来突出重点词语的作用,增强文本表示,最后通过全连接神经网络分类器进行分类,输出标准化的结果。

9） 杜等提出了一种基于多尺度残差图卷积网络的临床术语标准化方法,该方法采用多尺度残差网络来捕获临床文本的不同长度的文本模式,并基于图卷积神经网络抽取标签之间的层次关系,以加强标准化的能力。

1. 算法分析

根据上述归纳总结，我们将逐一分析以下，采用基于人工和规则、机器学习以及深度神经网络的方法的优缺点。

1. 采用基于人工和规则的临床术语标准化方法

采用人工和规则的临床术语标准化方法，虽然有了人力的加入，在准确率方面有所提高，但是相较于前两者的速度而言，还是不能够称为一个良好的算法。对于一个待预测的样本集，其时间复杂度几乎是O(N^2)。并且人工进行标准化，需要具有大量医疗知识的医护人员进行，这无疑是对社会资源的一种浪费。人工进行，同时也会因为个人的主观因素，导致分类错误，这同样也是不容忽视的一点。

1. 采用机器学习的临床术语标准化方法

采用机器学习的方法进行临床术语标准化，其优点是在速率上大大提高，并且也有大量的python包，例如sklearn等，使得调用机器学习也变得方便起来，但是其缺点也不容小觑，机器学习的方法，例如KNN,Bayes等方法，只能够捕捉语句表面语义，其内在的更深层次的信息则无法捕获。这使得其在面对更加相近的语义时，错误率会大大增加。

1. 采用深度神经网络的临床术语标准化方法

采用深度神经网络，其优点在于神经网络在文本建模上具有强大的表征能力，不仅仅可以更好的表示词语和文本，还可以学习到词语上下文之间的关系，以及重要词语的信息。但是其缺点在于，处理文本信息的速度实在过慢。如今主流方法是采用BERT进行语义标准化，但是BERT模型具有大量参数，这不仅是其优点，同时也是其缺点，使得每一次比较都需要耗费大量的时间。

综上所述，在实现难度方面，从小到大，依次是机器学习，深度神经网络，人工和规则。从可行性角度来讲，为了达到一个良好的准确度，其准确率从大到小，依次是，人工和规则，深度神经网络，机器学习。

1. 拟使用算法

根据上述调研反映出的现有模型的缺陷，本项目针对此痛点，本项目采用传统统计学习方法和深度神经网络相结合，通过采用BM25算法和SimCSE方法相结合的方法进行术语标准化。当然，该方法也存在一定缺陷，这一点将在之后介绍。

1. BM25算法

BM25是一种用来评价搜索词和文档之间相关性的算法，它是一种基于概率检索模型提出的算法。其计算方法，简单描述为当我们有一个query和一批文档Ds，现在要计算query和每篇文档D之间的相关性分数，我们的做法是，对query进行语素解析，生成词元qi；然后，对于每个搜索结果D，计算每个语素qi与D的相关性得分，最后，将qi相对于D的相关性得分进行加权求和，从而得到Query与D的相关性得分。

其一般公式为

图片包含 图形用户界面

描述已自动生成

其中Wi为表示单词权重，即IDF

图片包含 文本

描述已自动生成

Ri则表示词元qi与文档d的相关性得分

文本, 信件

描述已自动生成

1. SimCSE算法

对于语义相似度而言，其最重要部分则是在于句向量表征。句向量表征技术一直都是NLP领域的热门话题，在BERT前时代，一般都采用word2vec训练出的word-embedding结合pooling策略进行句向量表征，或者在有训练数据情况下，采用TextCNN/BiLSTM结合Siamese network策略进行句向量表征。在BERT时代，人们借助预训练语言模型的固有优势，一般采用BERT模型的[CLS]向量作为句向量表征。2019年，SBert诠释了在存在训练数据的情况下，如何在BERT模型上获取更好的句向量表征。2020年出现的BERT-flow和BERT-whitening主要是对无监督语义匹配的改进，而有监督语义匹配的SOTA一直是SBERT。直到2021年，SimCSE论文的发表，该模型引入对比学习的思想，大幅刷新了有监督和无监督语义匹配SOTA，更让人惊叹的是，无监督SimCSE的表现在STS基准任务上甚至超越了包括SBERT在内的所有有监督模型。SimCSE性能如图1所示。

一些文字和图片的手机截图

描述已自动生成

图1 SimCSE与其他模型对比

虽然无监督模型或许才是该论文的主要部分，但由于篇幅问题和项目数据集因素，我们在这里仅仅讨论SimCSE有监督学习部分，无监督部分不过多讨论。

1. 对比学习

对比学习的思想是去拉近相似的样本，推开不相似的样本，而目标是要从样本中学习到一个好的语义表示空间。一个良好的对比学习方法，应当具备Alignment和Uniformity。

Alignment指的是相似的例子，也就是正例，映射到单位超球面后，应该有接近的特征，也即是说，在超球面上距离比较近；

Uniformity，指的是系统应该倾向在特征里保留尽可能多的信息，这等价于使得映射到单位超球面的特征，尽可能均匀地分布在球面上，分布得越均匀，意味着保留的信息越充分。分布均匀意味着两两有差异，也意味着各自保有独有信息，这代表信息保留充分。

1. 在SimCSE中，我们构造这样一个三元组（X，X+，X-），将多个三元组同时输入到文本向量抽取的模型中进行特征抽取，既然是对比学习，则需要正样例与负样例，在SimCSE中其中X+作为X的正样例，X-以及同一batch中不同语句作为负样例。构造方法如图2所示

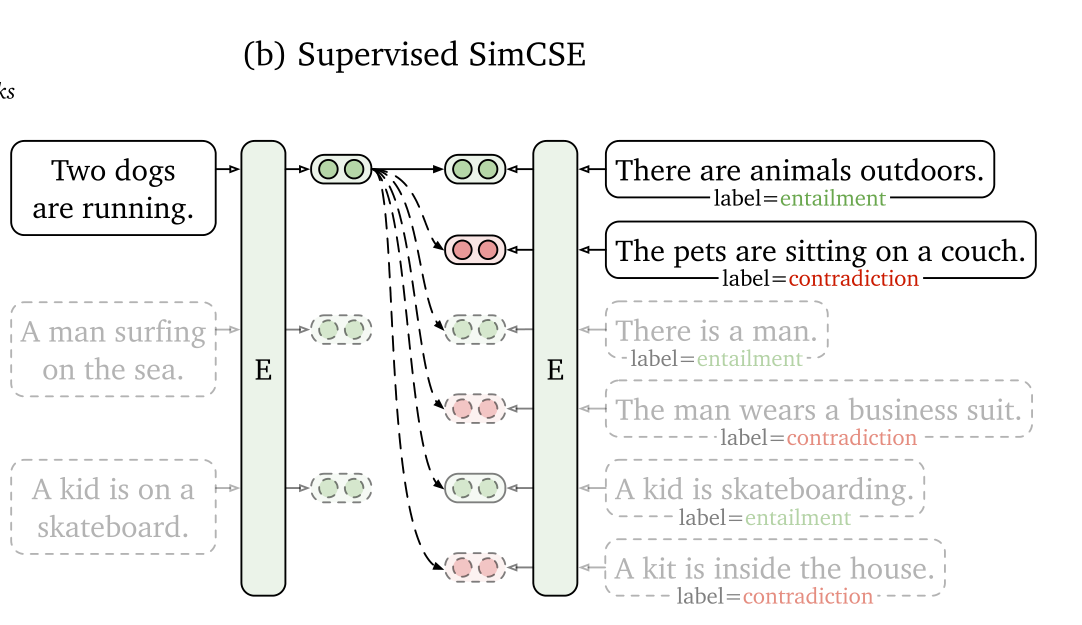


图2 SimCSE正负样例构造方法

其损失函数本质上仍然是采用infoNCE loss（info Noise Contrastive Estimation loss），主要的修改则在于将infoNCE loss中点乘操作换成了余弦相似度。分子是正例对的相似度，分母是正例对+所有负例对的相似度，最小化infoNCE loss，就是去最大化分子的同时最小化分母，也就是最大化正例对的相似度，最小化负例对的相似度。

图片包含 文本

描述已自动生成

1. 主要步骤介绍
2. 数据预处理

经过我们对于数据集的初步分析，数据集中存在着一些问题，不利于语义相似度的分析。例如，存在英文专业术语，罗马数字以及意图不明等描述不统一，对于之后的工作，无论是BM25还是SimCSE都有一定的影响。因此，我们借鉴了ark-nlp的工作，对于数据进行预处理，将我们能够筛选出来的描述进行了统一，其中包括了罗马数字替换，英文术语翻译，歧义语句互换。

1. BM25粗筛

当前临床术语标准化主要是和国际疾病分类 ICD-10 北京临床版v601标准数据集进行比较，那么其必然会存在一个问题，如果我们使用深度神经网络，则要和标准数据集中的每一项都进行比较。据分析，标准数据集一共存在4万多项术语，假设每一次匹配都需要1s，那么对于一个待标准化的样例来说，其需要耗费的时间则需要耗费4万多秒，即11个小时。并且，这个时间随着医疗水平的发展，各个病症的不断细分，时间还会继续增加。这无论对于临床医学术后诊断，还是病历立即处理都是不可接受的。仅仅两个病历，都需要耗费快一天的时间来处理。

因此，针对此问题，我们采用BM25算法进行初步筛选。临床医学虽然数量较多，但是不同类型手术还是存在着较大差别，所以我们利用该特性，对于每一次的样例进行筛选，筛选出部分与之相近的答案，之后再进行细筛，从此减少时间耗费，提高标准化速率。

虽然以BM25来检测文本相似度仍需要一定时间，但是其时间消耗已经在可以接受的范围之内，当然也可以采用更加快速的算法进行筛选，例如TF-IDF,Jaccard系数，基于MED的相似度算法算法

1. SimCSE细筛

train：

SimCSE发表在EMNLP 2021，其取得的结果是压倒性的，在这里我们主要是使用其有监督版本，我们以三元组的形式进行输入，当然也可以直接采用二元组的形式进行输入，最后结果估计不会存在太大变化。

predict：

在得到最终的训练模型之后，我们对于一个输入样例进行encode，以及BM25得到的初步筛选结果进行encode，最后比较二者相似性。当然，不可否认，这里处理方式忽略了两个词语之间的内在特征联系，存在在一定的片面性。

1. 代码实现与解释性标注

见附带文件。

1. 项目总结评估及未来发展展望
2. 代码评估

这次模型采用BM25粗筛+SimCSE细筛的方法，将传统方法和深度神经网络结合起来，分别发挥他们的优势并相结合。最终在单个相似度检测的任务上，其准确率达到了94%以上。但是模型仍然存在着一些问题，经过SinCSE进行编码过后的向量，如果采用余弦相似度直接进行比较，缺少深层次的信息交互，尽管速度相比之下会快很多，但是其深层次的特征并没有被很好的挖掘出来，从而导致准确率降低。

1. 项目总结

通过这次项目，我明白了有时候深度神经网络在实际应用环境下，其实并不能算做一个最优解。虽然其准确率高，但其所耗费的时间太多，往往很难在现实生活中应用，这也是如今人工智能难以落地实现的主要原因。并且，在代码设计阶段，应当考虑到来自各方面的因素，确保各个模块连接稳定，不要当你想要把整个模型链接在一起的时候，发现一个模块的输出和另外一个模块的输入其实并不对应。最后，对于数据的分析有时候远要比对模型的设计来的重要。一个模型最后达到的效果必然和数据集噪声大小、质量高低有着密切的联系。就像最近火热的prompt learning中的CLIP，采用4亿图文对，加以清洗，才能够达到那么经验的效果。

1. 未来发展分析

未来对于临床术语标准化的发展，我觉得最终会走向prompt learning的道路，或者说是自然语言处理最终会走向prompt learning。如今的与训练模型提供了一个非常好的初始化参数，针对于下游任务我们需要在这组参数的基础上进行pre-train，当我们使用了prompt learning之后，我们可以将其过程整理为“Pre-train, Prompt, and Predict”，我们可以在每一个文本之后加上适当的提示信息，例如，病症是’[mask]’。当然这只是个例子，具体的提示模板还需要进一步研究。