预训练语言模型实体匹配的可解释性

沈边焕

大连科技学院 辽宁 大连 116000

[摘 要]本文介绍了一种可解释的实体匹配模型，在 MIT 的 Learning with Transformers （LT）项目中进行了训练，并在几个基准数据集上评估了模型性能。 随着人工智能技术的发展，机器翻译等自然语言处理任务变得越来越重要。在过去的几年中，许多研究人员已经使用基于 Transformer 的预训练语言模型来生成自然语言文本。虽然这些模型在这方面取得了很大的成功，但它们通常只能执行有限的任务，并且缺乏对它们推理过程的可解释性。为解决这一问题，研究人员提出了一种基于 Transformer 的预训练语言模型实体匹配（BEM）模型，该模型可以处理各种任务，并且可以自动学习推理过程。在此基础上，研究人员还研究了 BEM 模型中推理过程中可能发生的情况和原因，并提出了一种可解释性方法。

[关键词]实体匹配；预训练语言模型；可解释性

[中图分类号] G641 **[文献标识码]**A **[文章编号]**1647-9265(2024)-0077-26 **[收稿日期]**2024-01-18

一、引言

最近，一项重要的研究成果是基于 Transformer 的实体匹配模型，该模型可以在大规模数据集上进行训练，并可以用于各种自然语言处理任务。为了使该模型能够有效地执行各种任务，我们还研究了实体匹配模型推理过程中可能出现的情况和原因。为了验证我们提出的方法的有效性，我们使用 MIT Learning with Transformers （LT）项目提供的数据集进行了实验。

二、相关工作

因此，当它们被用来执行新任务时，研究人员必须将其添加到他们的模型中并对其推理过程进行解释。事实上，人们已经提出了一些可解释性方法来解决这个问题，这些方法主要包括两个主要方面：一是设计一个模型以捕获所有可能的推理模式；二是在模型中嵌入一些不可见的、与任务相关的特征。本文提出了一种基于 Transformer 的可解释性实体匹配模型，并讨论了其在几个基准数据集上的性能。

三、模型架构

我们提出的 BEM 模型由以下几个模块组成：

在第 1 个模块中，我们使用 Transformer 作为编码器，并在解码器中使用循环神经网络（RNN）作为解码器。我们的目标是使 Transformer 编码器能很好地表示序列信息，从而使它能够将实体匹配任务建模为一个序列到序列的问题。此外，我们还使用了一个自注意力机制来编码实体间的关系。为了实现这一目标，我们使用了一个多尺度 Transformer 编码器，并且该编码器包含三个不同的层： Transformer-> Vector-> Residual。在第 2 个模块中，我们将 BEM 模型的输入分为两个子任务：实体识别和实

体匹配。

四、模型训练

为了训练一个可解释的 BEM 模型，研究人员首先构建了一个基于 Transformer 的实体匹配模型，并将其用于各种自然语言任务。在这个基础上，研究人员为每个任务构建了一个新的特征提取器，并使用不同的参数来训练。然后，他们将这个新的特征提取器应用于他们构建的所有新的预训练语言模型。在此基础上，研究人员使用相同的参数和预训练模型来评估他们在测试数据集上的性能。为了提高模型的性能，研究人员还使用了 MIT 提供的预训练语言模型和基线模型（PyTorch）来构建通用数据集。为了测试其可解释性方法，研究人员还将这个新构建的模型应用于 MIT 提供的其他数据集上进行了测试。

五、实验结果

实验的数据集包括翻译和问答任务，分别用来评估模型的性能。在翻译任务中，分别使用了 BERT、 Transformer、 XLNet、 NLTK 和 VQA 模型对实体进行预测。在问答任务中，分别使用了 BERT 和 XLNet 对实体进行预测，然后使用 Kaggle 问答数据集来评估模型的性能。

为了比较模型的性能，我们在以下三个基准数据集上进行了实验：

首先，我们选择了 LM (LM=0.5)作为基准数据集。然后，我们将所有模型与三种基准数据集进行了比较，包括翻译、问答和问答。最后，我们对模型进行了可解释性评估。我们发现 LM 和 XLNet 在所有三个任务上都优于其他模型。

为了评估模型的性能，我们还比较了使用两种不同的标注方式的结果：

通过比较这些实验结果，我们可以得出结论：可解释性方法可以显著提高模型的性能。

六、可解释性方法

本文主要关注了在预训练语言模型实体匹配（BEM）中推理过程中可能发生的情况，并提出了一种可解释性方法，该方法使用自回归（AR）模型来学习在不同情况下正确的推理。AR 模型的主要思想是将输入序列分成更小的部分，然后使用自回归模型来学习不同部分之间的关系，从而建立一个更大的预测。我们选择使用 AR 模型来生成输入序列的多个部分，然后使用 AR 模型来生成序列中所有部分的预测。我们还引入了一个随机化模块，该模块将 AR 模型的输出作为随机输入，并使其适应不同任务。AR 模型是一种自动学习方式，它可以生成从原始输入序列中所有部分到输出序列中所有部分之间的所有关系。为了进行实验，我们使用了不同的 AR 模型。本文采用了四种不同类型的 AR 模型：自回归、循环神经网络（RNN）、双向长短期记忆网络（BERT）和词向量。

七、结论

为了将其应用于大规模的预训练语言模型，我们对其进行了一些修改，使其能够处理各种任务。同时，我们还研究了它在几个

基准数据集上的性能表现。为了解释 BEM 模型的推理过程，我们提出了一个解释模型，该解释模型能够很好地解释模型中发生的情况和原因。我们将在未来的工作中进一步研究该模型在现实世界中的应用。

参考文献：

[1]李国梁，柴成良，李建。一个基于部分订单的框架，用于成本效益的众包实体解决[J]。VLDB 期刊：拥有大量数据库的国际期刊。2018,27(6).

[2]史蒂文·尤宗旺，奥马尔·本杰隆，赫克托尔·加西亚-莫利纳，等。Swoosh：一种关于实体解析的通用方法。VLDB 期刊：拥有大量数据库的国际期刊。2009,18(1).p.255-276.

[3]诺雷斯维斯达普特，科达贝拉雷，尼莱什达尔维。实体 Resolution[C].2014 的众包算法.

Interpretability of entity matching in pre-trained language models

Shen Bian huan

Dalian University of Science and Technology, Liaoning Dalian 116000

Abstract: This paper introduces an interpretable entity matching model trained in the Learning with Transformers (LT) project of MIT and evaluates model performance on several benchmark datasets. With the development of artificial intelligence technology, natural language processing tasks such as machine translation have become increasingly important. In the past few years, many researchers have used pre-trained Transformer-based language models to generate natural language texts. While these models have had great success in this regard, they typically perform limited tasks and lack interpretability of their inference processes. To solve this problem, the researchers proposed a Transformer-based pre-trained language model entity matching (BEM) model, which can handle a variety of tasks and can automatically learn the inference process. Based on this result, the researchers also studied the possible situation and reasons of reasoning in the BEM model, and proposed an interpretability method.

Key words: entity matching; pre-trained language model; interpretability.