面向自动问答系统的短文本相似度计算

黄 震*
HUANG Zhen

摘 要 基于文本的自动问答系统一般包含问句处理、问句匹配和答案选取三个模块，衡量系统的重要指标（包括答案选取的效率和准确性）与问句匹配模块的相关度最高，因此从一个自动问答系统落地应用的角度出发，本文在问句匹配模块中提出了一种有监督和无监督相结合的短文本相似度计算方案。实验部分设计了基于 Siamese Network 框架的短文本相似度计算与基于 word2vec 词向量的无监督计算方法对比，验证了有监督计算方法在准确率方面的优势。

关键词 短文本相似度计算；自动问答；Siamese Network

doi：10.3969/j.issn.1672-9528.2020.11.063

0 引言

短文本相似度是两个短文本语句相似程度的数值表示（一般在 0～1 之间），在自动问答系统中就是通过计算短文本的相似度来进行语句匹配。作为自然语言处理（NLP）的基本任务之一，短文本相似度计算也在业界的语言类应用中占有重要地位。在自动问答系统任务中，我们首先会创建问答数据库，存储问答任务中常用且描述准确的问题－答案对，这些存储的问题一般被叫作标准问题。当用户使用系统进行提问时，我们会将用户输入的问题文本与数据库中的标准问题进行相似度计算，找出与用户问题最相似的标准问题，返回与标准问题对应的答案给用户，从而实现自动问答功能[1]。短文本相似度计算方法主要有无监督的相似度计算和有监督的相似度计算两种。

1 短文本相似度

1.1 文本表示

在进行短文本相似度计算之前，首先要做的就是让计算机能够识别输入的文本内容，即文本表示。因为计算机不能对文本直接处理，所以需要将文本进行数值向量化转换。 在无监督相似度计算方法中，文本表示的主要方式有基于统计的词频－逆文档频率（TF-IDF）方法，根据词语的 TF-IDF 值来衡量其在句子中的重要程度并以此赋予权重进行向量化表示；基于词嵌入（word embedding）模型的方法，从原始文本或外部语料库中学习词嵌入，然后将词向量通过不同的加权算法转换为句向量。

1.2 相似度计算

无监督的相似度计算通过不同的文本表示获得短文本的句子向量表示，之后对两个短文本的句子向量进行距离度量，（认为向量间距离越近就越相似）获得短文本相似度。常见距离度量方法有余弦距离、欧式距离和曼哈顿距离等[2]。

有监督的相似度计算是先对问题语料进行标注，标注两个短文本是否相似，然后根据标注的数据进行深度学习建模，将短文本相似度计算转换成一个类型判定问题，通过模型的端到端学习，直接求解出短文本的相似度值[3]。目前比较经典的方法是基于孪生网络（Siamese Network）框架，如图 1 所示。

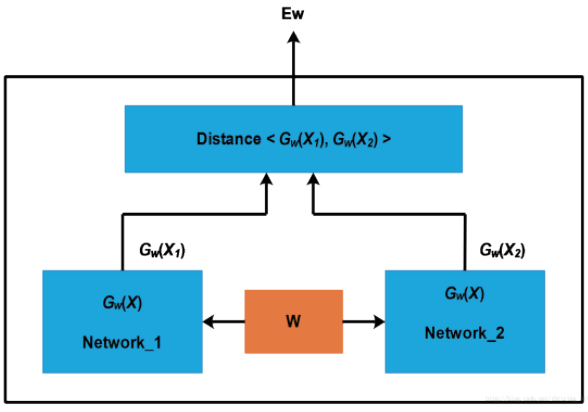

图 1 Siamese Network 框架图

Siamese Network 有两个结构相同，且共享权值的子网络，分别接收两个短文本输入 X1 与 X2，将其转换为向量 Gw(X1) 与 Gw(X2)，再通过某种距离度量的方式计算两个输出

* 东南大学软件学院 江苏南京 211189

向量的距离 Ew，最终获取短文本的相似度值。子网络的具体实现可以选择 CNN、LSTM 等，训练 Siamese Network 采用的标注样本是一个 tuple (X1,X2,y)，标签 y=0 表示 X1 与 X2 属于不同类型（不相似、不重复），y=1 则表示 X2 与 X2 属于相同类型（相似）。

2 实验

2.1 实验数据与实验环境

本文实现了有标注数据的基于 Siamese Network 的短文本相似度计算，数据来源于 CCKS 2018 微众银行客户问句匹配大赛，数据集大小为 100000 条，标签代表文本是否相似，其中相似与不相似的比例为 1:1。数据标注示例如下表 1 所示。

表 1 标注数据表

句子 1	句子 2	标签
第一次使用，额度多少	我的额度多少钱	1
借款需要什么证明	不给开还发什么消息	0

实验采用的开发语言为 Python3，选择了基于 TensorFlow 的一个深度学习框架 Keras。

2.2 实验设计

本文实现了基于 Siamese Network 框架的有监督的短文本相似度计算模型和基于 word2vec 词向量的无监督计算模型作简单的对比实验。有监督的计算方法采用典型的 Siamese Network 框架，计算模型如图 2 所示。

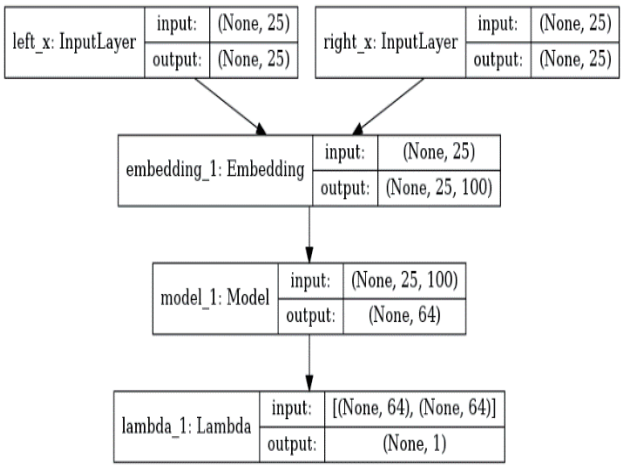

图 2 Siamese Network 框架计算模型

两个短文本语句分成左右两个部分进行输入，编码层权重共享，使用两个双向 LSTM 进行编码，最后基于曼哈顿空间距离计算两个短文本语义表示相似度 [4]。

无监督的计算方法使用 gensim 包从数据集中学习词嵌入，然后将词向量通过简单的加权取平均的方式转换为句向量，最后基于余弦距离计算两个短文本的文本相似度。

2.3 实验结果分析

表 2 实验结果表

模型	训练集	测试集	训练集准确率	测试集准确率
有监督	80000	20000	0.8254	0.8062
无监督	80000	20000	0.6014	0.5688

无监督的短文本相似度计算与有监督的计算方法相比有一定的优势：不需要花费大量的人力资源对数据进行标注，同时在时间复杂度和适用性方面表现更好，但是在准确率方面，有监督的计算方法表现更优秀。无监督计算的弊端在于文本的句向量生成是通过词向量人为地加权求和，没有涉及句子句型、语法，结构等，难以包含句子的整体语义信息，并不能很好地表示文本。本文的实验也验证了有监督的相似度计算模型有着更好的准确率。

3 总结与展望

本文在面向自动问答系统应用的角度上提出了一种有监督和无监督相结合的短文本相似度计算方案：在问答系统启动阶段将问答数据库中的标准问题经过孪生网络生成短文本的句向量并将其存储，当系统接收用户的问题输入时，只需求解该问题的句向量，然后与数据库中存储的标准问句向量进行距离度量，根据相似度的大小排序就可以获得最相似的标准问题，从而返回给用户较为准确的答案。另外，可以将 BERT 模型代替原来孪生网络中的双向 LSTM，获取具有更多语义信息的句向量，下一步将在此方向上进行研究。

参考文献：

[1] 李月，周江．一种基于文本相似计算的校园智能问答系统设计 [J]. 现代信息科技，2019,3(22):9-12+17.

[2] 吴佐平，刘迪，张千福，等．面向客服的自动问答系统的相似度计算研究 [J]. 信息技术，2020,44(3):99-103.

[3] 郭浩，许伟，卢凯，等．基于 CNN 和 BiLSTM 的短文本相似度计算方法 [J]. 信息技术与网络安全，2019(6):61-64.

[4] Neculoiu P，Versteegh M，Rotaru M．Learning Text Similarity with Siamese Recurrent Networks[C]// Repl4NLP workshop at ACL 2016. 2016.

（收稿日期：2020-09-26 修回日期：2020-10-19）