“智慧政务”中的文本挖掘应用

摘要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

对于问题 1,通过 Positioned 对城市建设和市政管理表、留言表进行去重,得到不重复的分类信息。利用 jieba 中文分词工具对留言信息进行分词,并通过 TF-IDF 算法提取每个留言描述的前 5 个关键词。再利用 TF-IDF 算法得到每个留言描述的 TF-IDF 权重向量,采用 K-means 对 TF-IDF 权重向量进行聚类,得到 7 个质心。分别求出距离各个质心最近的 5 个留言并且分类,结合城市建设和市政管理表的 PositionFirstType 字段,根据 F-Score 算法,进行统计分析，对分类方法进行评价后建立一级标签模型。

对于问题 2,根据数据挖掘与分析的职位特征，将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标。利用发散性思维,再分别对筛选出来的结果按照热点问题、热点留言明细两个方面对其进行多方面系统地统计,最后制作成图表。

对于问题 3,根据研究结果,通过分析目前的留言热点等问题,给相关部门的答复意见提出可行性的建议。

Application of text mining in "intelligent government affairs"

Summary

In recent years, with WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly depends on artificial to divide and hot spots of relevant departments work has brought great challenge. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government.

For question 1, through Positioned on urban construction and municipal management table, message table to heavy, don't repeat the classification of the information. Jieba Chinese word segmentation tool is used to segment the message information, and the first five keywords of each message description are extracted by tf-idf algorithm. Tf-idf algorithm is used to obtain the weight vector of tf-idf for each comment description, and k-means is used to cluster the weight vector of tf-idf to obtain 7 centroid. The 5 comments closest to each center of mass were calculated and classified respectively. Combined with the PositionFirstType field of the urban construction and municipal management table, statistical analysis was conducted according to the f-score algorithm, and a first-level label model was established after the evaluation of the classification method.

For question 2, according to the job characteristics of data mining and analysis, the comments reflecting the problems of specific places or people in a certain period of time are classified, and a reasonable heat evaluation index is defined. Using divergent thinking, and then the selected results in accordance with the hot issues, hot message details of the two aspects of its many aspects of systematic statistics, finally made into a chart.

For question 3, according to the research results, by analyzing the current hot topic of comments and other issues, put forward feasible Suggestions to the replies of relevant departments.

目录

一、 简介

1.1 挖掘意义

关于阅读，相信大家都不陌生，在学习方面，我们接受的传统语文教育中阅读理解是非常常规的考试内容；在生活中，我们也常常会遇到阅读文件、论文、书籍等应用场景。除去欣赏优美的语言艺术，更多情况下，我们只是需要从文本中查找某一些片段来解决我们的实际问题。比如，通过查找法律文献中的一些段落来解决我们的法律疑惑，这时并不需要精读整个法律文献；就算对于小说，有时候我们也只是想知道其中一些特殊细节，并不想花时间去通读整个小说。但是，这对人类来说无疑是一个难题。浩如烟海的书籍作为古往今来人类智慧的结晶，其中内容往往很难在短时间内为人掌握，就算是略读也不容易，更不必说从浩繁卷帙中准确的定位答案。因此，我们希望智能阅读技术能够在这方面提供一些帮助。

为了构建智能文本挖掘模型，学界对于机器阅读理解的研究从未止步。机器阅读理解作为目前热门的自然语言处理任务，目标是使机器在能够理解原文的基础上，正确回答与原文相关的问题。提高机器对语言的理解能力。机器阅读理解技术的发展对信息检索、问答系统、机器翻译等自然语言处理研究任务有积极作用，同时也能够直接改善搜索引擎、智能助手等产品的用户体验，因此，以阅读理解、文本挖掘为契机研究机器理解语言的技术，具有重要的研究与应用价值。

1.2 挖掘目标

我们要构建一个智能的文本挖掘模型。模型可以起到辅助阅读的作用，帮助人们显著提高阅读的效率。具体到使用情景上，对于用户输入的问题与文档，模型可以定位到文档中能帮我们回答问题的所在行，或者直接给出明确的某个词作为答案输出。

更形式化地，我们要解决的智能阅读模型的构建问题可以被解释为一个三元组<D,Q,A>。三元组由文档 D，问题 Q 和答案 A 的答案组成。作为最大粒度的输出要求，我们需要把答案 A 定位到文档中的某一句话，为了解决这个问题，需要同时利用问题 Q 中的信息与文档 D 中的全部上下文信息。

1.3 挖掘流程

如图 1 挖掘主要分为两大部分，预处理部分和候选答案评分部分。其中预处理包括分词，去停用词，将词组向量化（word2vec）。候选答案评分部分为核心步骤，为了进一步挖掘上下文信息，将预处理得到的词向量放入 Bi-LSTM 网络,为了进一步凸显处答句整句的语义信息，对答句进行 sentence2vec,最后使用注意力机制对 Bi-LSTM 的解码端进行信息整合。最后对已经有了评分的候选集，设置阈值，输出预测结果。

2 . 1 word2vec

为了将语料输入神经网络进行训练，我们首先要将自然语言符号表示成计算机能够理解的数字形式。

一个自然的想法是把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。这就是独热编码形式（One-Hot）。独热编码虽然方便易懂，但也有显而易见的不足：首先 One-hot 编码的维数由词典长度而定，过于稀疏，存在降维难问题，给计算造成了很大不便；其次， One-hot 编码下任意两个词之间都是孤立的，丢失了语言中的词义关系。

Word2vec 是 Mikolov 在 2013 年提出的用于快速有效地训练词向量的模型[3]。作者的目标是要从海量的文档数据中学习高质量的词向量,该词向量在语义和句法上都有很好地表现,已经广泛应用于自然语言处理的各种任务中。 Word2vec 包含了两种训练模型,分别是 CBOW 和 Skip_gram 模型,如图 2 所示。中 CBOW 模型利用上下文预测当前词,而 Skip_gram 模型利用当前词预测其上下文。

三、 对回答候选集评分

3.1 Bi-LSTM 层

为了尽可能保持语句中词组之间的上下文联系，我们决定将通过上一步 word2vec 得到的词向量放入循环神经网。同时为了获得尽可能多的上下文记忆信息，我们最终选择了 Bi-LSTM 这一模型。

3.1.1RNN 和 LSTM

循环神经网络（Recurrent Neural Network,RNN[5]）近年来由于其良好的性能代替深度神经网络（Deep Neural Network,DNN）成为主流自然语言处理建模方案，相对于 DNN，RNN 在隐层上增加了一个反馈，即 RNN 隐层的输入有一部分是前一级的隐层输出，这使 RNN 能够通过循环反馈看到当前时刻之前的信息，赋予了 RNN 记忆功能，能较好的表征上下文的语义。这些特点使得 RNN 非常适合用于对自然语言进行建模。如图 3 所示，所有的 RNN 都具有一种重复神经网络模块的链式形式,在标准 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。

RNN 虽然看似简单，却也有一些严重的缺点。首先是过拟合，在现在广泛应用的端到端 RNN 系统中， RNN 对上下文相关性的拟合较强，而这会导致 RNN 相比与 DNN 而言更易出现过拟合问题。其次是梯度消失和梯度爆炸，因为 RNN 比 DNN 更复杂，海量数据环境下的 RNN 模型训练难度较大，容易出现梯度消失和梯度爆炸问题，导致在构建大型系统方面表现较差。第三是 RNN 对较长的问句有严重语义遗忘问题，当相关信息和当前预测位置之间的间隔不断增大时， RNN 有限的记忆能力会因为信息不断增多而耗尽，因此当间隔增大，一部分初始的记忆信息就会被遗忘，这些长期依赖的遗忘和缺失最终会导致问句语义丢失。为了避免以上问题，我们选择用 LSTM[6]这一 RNN 变种。

3.2 总结

本节中所提出的神经网络架构如图 7 所示: s 中每个元素的值，即其对应位置的答句与问句的匹配分数，也可以理解为该答句是正确结果的可能性 P。

四、 实验评估

本节中我们将对上文中我们建立的模型进行实验验证，通过与传统方法以及其他深度学习方法进行比较分析，从而说明我们模型的合理性与有效性，同时介绍与分析模型的最优参数调整过程与各类参数对模型性能的具体影响。

4.1 实验平台

实验环境的软件硬件配置如表 2 所示:

使用到的 python 库有:

• tensorfow[9]:由 Google 大脑小组的研究员和工程师们开发出来开源机器学习框架 tensorfow,我们应用了其中 Bi-LSTM 、 Attention 、 Word2Vec,Softmax 这些深度学习算法。

4.2 相关结果

(1) 在不同数据集上验证模型的性能
如图 9 所示，为模型在题目给定数据集，百度 WebQA 和保险行业语聊库上的相关性能。在三个数据集上，我们的模型均表现出了优异的性能，说明该模型可以成功捕捉到不同语料库中的问答关系，具有较强的泛化能力。

(2) 使用不同的词嵌入方法的相关性能
在实验过程中，为了确定最佳方案，我们分别使用 Skip_gram 算法， CBOW 算法与传统的 One-hot 编码算法对语料进行词嵌入处理，并将测得性能进行了对比。实验结果，如图 10 所示。 Skip_gram 与 CBOW 同属 W ord2vec 方法，而使用 Word2vec 获得的 F1-score 是只使用 One -Hot 方法的 3.5 倍，这说明 Word2vec 算法的性能要明显优于 One-hot 编码，而 Word2vec 中， Skip_gram 与 CBOW 性能相近，多次实验统计结果的平均水平表明， Skip_gram 的 F1-score 比 CBOW 高 6.484%， ACC 低 1.337%， MAP 高 1.639%， MRR 高 1.467%， Skip_gram 略好于 CBOW，因此我们最终采用 Skip_gram 算法进行词嵌入。

(3) 采用 sentence2vec、普通 RNN 生成句向量和不采用句向量下相关性能

如图 11 所示，采用 sentence2vec ， F1-score 比基于 RNN 生成句向量的模型高 15.405%，比不使用句向量的模型高 19.776%。经过分析，与基于 RNN 生成句向量的模型相比，sentence2vec 不会出现梯度消失问题，且训练速度明显加快；而 sentence2vec 嵌入的句向量兼顾了上下文提取句意，有助于最终将答案定位到具体的句子；与不使用句向量、仅用词向量的模型相比，性能提升很大。

(4) 采用 Attention-over-Attention 机制和与不采用机制下相关性能

如图 12 所示，采用 Attention-over-Attention 机制后， F1-score 比起单一注意力提高了 22.350%。 Attention-over-Attention 机制捕捉了答案到问题、问题到答案的双重注意力，使模型可以利用问句和答句之间的交互信息，这样的机制比起只依赖单一注意力的普通模型，兼顾利用的信息更多，因此性能更优。

(5)本文模型与其他模型相比较

如图 13 所示，我们在出题方所给的数据库上，我们分别对比了基于传统方法进行语义分析的模型与基于卷积神经网络的模型，我们的模型均优于上述模型。

5.1 未来改进

由以上论述可知，对于输入的问题，我们的模型可以出色地完成将答案定位到所在句子的任务。下一步，我们希望继续改进模型，实现更细粒度的答案提取。如果需要将答案定位到词语，我们可以去掉嵌入句向量的步骤，用 Bi-LSTM 得到的蕴含上下文信息的词向量进行 Attention-over-Attention 的计算。按列计算的注意力代表了问题中每个词选择答案的重要性，按行计算的注意力代表了文档中的每个词响应问题的重要性。正反注意力经过点乘，最终获取的注意力向量每个分量值含义为文档中每个词与问题的匹配分数。统计每个词对应分量值之和，即以该词作为答案时与问题的匹配分数之和，获得所有和值并排序，分数越高则该词是正确答案的可能性越大。