利用深度学习进行人机对话

梅 桐(北京航空航天大学实验学校(中学部),北京市 100000)

【摘 要】人机对话一直都是人工智能领域的核心问题,人机对话能够有效的组织信息,弥补了搜索引擎的不足之处。本文主要介绍了人机对话的背景,国内外的研究现状,以及针对人机对话的数据集。本文重点对人机对话中词转换成计算机能够理解的词向量以及人机对话的方法,评价标准做了重点的介绍。

【关键词】人机对话;深度学习
【中图分类号】TP181 【文献标识码】A 【文章编号】1006-4222(2018)08-0220-02

1 概 要

1.1 问题的研究背景以及意义

最近几年随着互联网的发展,尤其是社交网络的广泛普及,人们从互联网上获取到的信息也越来越多。语言信息又是其中最直接的一种,如何从众多的语言信息中发现对我们来说比较重要的信息便显得尤为重要。但是,目前的搜索系统像是百度,谷歌,必应等都没有很好的解决这个问题,很多时候搜索不到我们想要的答案。现在存在的搜索引擎大都是根据关键字来进行检索任务,这样的检索会存在着如下的问题:

(1)获取到的信息太多。当我们通过关键词进行搜索的时候,搜索引擎会给我们非常多的出现这些关键词的信息,但是这些信息是没有经过筛选过的,也就意味着大部分的信息是我们不需要的,还需要我们在进一步的甄别其中有用的信息。

(2)检索信息比较复杂。有时候我们所需要的信息并不是简单的通过几个关键词就可以找到具体的内容,用户有时候也不能很好的去表述自己的意图,那么搜索引擎也无法帮助我们找到我们想要的信息。

显然,传统的搜索引擎并不能很好的帮我们解决搜索信息的问题,我们更加需要的是通过自然语言的形式来表达我们想要的问题。那么,自然语言对话就出现了。尤其是深度学习技术的发展,使得人类与机器之间的对话变得越来越容易实现。想象一下在未来的某一天,你可以和虚拟的个人助手交谈工作,你不需要在屏幕或者触控板上点击,只需要说出你的需要,你的虚拟助手就可以迅速的帮你实现。当你下班后,虽然家中一个人都没有,但是你还有一个可以与之谈心的情感聊天机器人。因此,对话系统是一个比较有发展前景的方向。

1.2 研究现状

自然语言对话一直都是人工智能领域最具有挑战性的任务,主要包括自然语言理解,自然语言之间的推理以及一些常识性知识的应用。从 20 世纪 60 年代开始,人们就开始思考如何让计算机像人类一样去理解语言,像人类一样进行思考。Start 是第一个关于自然语言理解的问答系统,由麻省理工大学在 1993 年发布,能够回答一些关于地理,历史等的常识性问题。AskHeeves 由密歇根大学设计,能够回答更多的问题,同时支持多语种的选择,但是给出的并不是问题的答案,而是一些网页,这跟我们的需求依然不是非常符合。这些工作大部分都是根据一些人类构造出的规则进行人机之间的对话,或者一些简短的问题回答。但是这样做往往有着非常明显的不足,比如人类去构造特征比较繁琐,而且大多数时候是不能发现完整的特征的。另外之前的相关应用都是特定场景下的,也就意味着只有很少的数据量以及模型的泛化能力非常之差。

最近几年随着硬件技术的进步,尤其是 GPU 的发展,使用深度学习的方法进行人机之间的对话也变得越来越容易,神经网络模型在解决序列到序列的问题的时候也有着天然的优势。比如,神经网络可以发现非常多的隐藏特征,这些特征往往是人类无法去认识到的,神经网络模型的泛化能力也比传统的方法要强很多。

2 数据集介绍

知乎是一个目前最为流行的中文的问答社区,提供了大量的高质量的问题以及问题的答案数据,并且这些数据是对公众开放的,非常容易获取的到,即使在你没有注册的情况下也可以获得绝大多数的问题以及大量的问题的答案,更为重要的是,问题的答案不仅仅只有一个,而是有很多个关于该问题的答案,并且可以按照点赞的数量对问题答案的质量进行排序,这样知乎就为我们提供了大量的训练数据以及测试数据。有一个很重要的问题就是,这些答案有的会非常的长,我们在进行问题的回答的时候,往往不想听到太长的答案,因此,我挑选出了其中答案质量比较高但是问题的回答也比较短的答案作为我们的标准答案,通常不多于 150 个字。

在构建这样一个数据集的过程中,首先是从知乎上获得了大量的问答的数据,然后对数据进行清洗,去除一些与问题关系性比较差的词,比如像“谢邀”,然后去除一些潜在的广告,最后挑出一些答案比较短同时回答的质量又比较高的词作为我们的数据。

最终,获得了大约 10 万条问题的描述以及相关的高质量的答案,将这一部分数据集的 60%作为训练集,20%作为交叉验证数据,20%作为测试集,共同组成整个算法的数据集。

3 将词转换成词向量

由于本文采用的网络结构为序列到序列的模型,因此如何将词表示成词向量也是一个非常重要的方面。由于我们是从知乎上获取的数据,因此我们的词表可能是非常大的,像是 one-hot 这种编码方式明显非常不适合这项任务,因为它表示一个词所需要的存储空间会随着词表的增大而增大,不适合这种词表非常巨大的词向量表示。通过多种词向量表示方法的对比,我最终选择了 Google 提出的 word2vec 的这种方法。

Word2vec 是一种基于概率的方法,主要有两种表示方式。一种是 skip-gram 方法,这种方法是根据当前的中心词去预测它周围的词的概率,通常是用当前的中心词与当前窗口大小中的所有词的相似度占此表中所有的词与当前中心词的相似度的比例,其中相似度采用两个词向量的点乘实现。另外一种

方法是 CBOW 算法，即连续词袋模型，这种方法是根据中心词周围窗口中出现的词的概率去预测当前的中心词的，速度上会有一定的提升。

4 网络的结构

对话系统的核心思想就是设计出一种表示方法，将输入的语言序列映射成一个输出的语言序列。在上一步中我们已经将语言序列编码成了一个词向量，那么每一个词有一个词向量表示，一句话就有一个更加长的向量来表示。算法主要包括两个过程，一个就是将我们已经转换好的词向量编码成一个特定长度的向量 L，这个向量中包含着我们编码的这句话的信息，然后我们在根据这个词向量去预测答句中出现的每一个词的概率，预测下一个词的时候不仅需要根据这个词向量，而且需要根据我们之前已经预测过的上一个词来共同参与预测的过程，概率我们用 softmax 来表示。

在机器翻译当中，L 是一种从源语言到目标语言的表示方法，但是在对话系统中，L 需要扮演的角色就更加的复杂了，因为对话的过程中可以产生的答案并不是一个，或者不像机器翻译那样非常的近似，因此 L 需要表示更多的特征，即可以非常丰富的表示从问句到答句的一些信息。

由于循环神经网络的概括和产生任意长度的输出序列的能力，在编码和解码的过程中，本文都是使用循环神经网络来做的。长短期记忆网络（LSTM）是一种非常特别的循环神经网络，LSTM 可以保存较长期间的行为，而不需要做非常多额外的操作，对话系统的任务中，往往有着比较多的比较长的句子，因此本文选择 LSTM 单元来作为循环神经网络的神经元。

4.1 编码的过程

在时间步 t 中，我们输入一个词的词向量 x_t 每一个隐藏神经元的计算按照如下的公式计算：

$$h_t=f(x_t,h_{t-1})$$

其中 h_t 表示当前神经元的输出，h_{t-1} 表示上一个时间步中神经元的输出，x_t 表示当前时间步内神经元的输入，然后我们用输入词序列的最后一个词输入后的神经元的输出（固定长度），来表示对这一整句话的概括，这个词向量表示整句话中的所有的信息。

4.2 解码的过程

在编码的阶段我们将输入的一句话编码成了一个固定长度的向量 x，那么我们在解码的过程中只要根据这个向量去求出每一个输出词的概率即可，第 t 个词的概率：

$$P(y_t|y_{t-1},\cdots,y_1,x)=g(y_{t-1},S_t,C_t)$$

其中 y_t 是输出词的 one-hot 表示，C_t 是上下文信息，g 是 softmax 函数，主要用于求概率，S_t 是第 t 个时间步的隐藏层神经元的输出，通过如下的方式计算：

$$S_t=f(y_{t-1},S_{t-1},C_t)$$

f 是非线性单元，由于 Relu 函数是目前最常用的激活函数，也由于其在反向传播的时候不会造成梯度消失等现象，本文选择了 Relu 函数作为激活函数。

4.3 模型改进

LSTM 和 GRU 都是设计来进行存储长期记忆的，实践证明它们都有着非常不错的效果，尤其是在一些与时序性比较强的数据中。但是 LSTM 有三个门，在反向传播的时候需要大量的计算，参数多需要的显存也比较大。GRU 创新性的将忘记门和输出门合成了一个更新门，显著的降低了参数的适量，加快了计算速度，在本文的试验中发现，GRU 和 LSTM 有着差不多的效果，但是明显的 GRU 的训练速度较快，因此改进的网络中，本文将编码和解码过程中的 RNN 用 GRU 来替代。

5 评价方法

之前的对话系统通常通过人工的方法去评价，即先建立一个测试集，然后与机器进行对话，人工判断对话是否更加的准确，但是这种方法需要大量的测试以及大量的人力资源，测试时间比较长。

DSTC2013 是第一个关于对话系统的公开比赛，有微软亚洲研究院，卡耐基梅陇大学以及本田研究院共同组织的一次比赛，主办方总共给出了 11 中评测指标，主要包括首位假设的准确率，L2-norm，以及 roc 等。本文最终选择 top3 准确率作为模型的评价标准。

首先需要计算模型给出的结果与标准结果之间的相似度，当结果相似度大于一个阈值的时候，认为模型的结果是正确的。然后从模型的多个预测结果中选出相似度最大的 3 个，然后与标准答案进行对比，只要有一个预测结果是正确的，那么我们认为该轮对话是正确的。

6 总结与展望

随着时代的发展，从海量的数据中发现有用的信息便显得尤为重要。人机对话是自然语言处理中的重要研究方向，深度学习的发展为人机对话提供了强大的技术支持，同时人机对话还可以减少人们的工作量，如银行柜台咨询人员，医院导诊人员等，都可以用相应的对话系统去替代，因此有着广泛的研究前景。

本文主要介绍了人机对话的背景，研究意义，当前的研究现状，所使用的数据集，以及进行对话使用的方法，评价标准等。

但是，由于本文的方法将对话系统中的编码过程编码成了一个固定维度的词向量，如果词向量的长度不够的话，那么很有可能漏掉很多的信息，如果太长的话，则会增加网络的训练时间，造成很多时间上的浪费。对于此次研究，仍有些不足之处，还需要加以改进。

参考文献

[1]宁长英.智能聊天机器人的关键技术研究.
[2]Oriol Vinyals.A Neural Conversational Model.
[3]A. Graves.Supervised Sequence Labelling with Recurrent Neural Networks.Studies in Computational Intelligence, Springer, 2012.
[4]Google Brain.Sequence to Sequence Learning with Neural Networks.
[5]Yoshua Bengio.Learning Phrase Representation using RNN Encoder-Decoder for Statistical Machine Translation.
[6]Junyoung Chung.Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

收稿日期：2018-7-12
作者简介：梅 桐（2001-），男，汉族，北京海淀人，高中在读，研究方向为深度学习。