中法工程师学院

研究生学位论文研究文献综述报告

**基于神经网络的语言模型的研究综述**

学科专业：工业工程

学生学号：ZY1624134

学生姓名：是黎彬

指导教师：荣文戈

完成日期：2017年12月9日

**摘要**

语言模型在自然语言处理领域起着非常重要的作用，该性能的好坏直接影响例如机器翻译和语音识别等很多任务的发展。随着循环神经网络的提出以及应用于语言模型上，计算机对人类语言的建模更加精确，但是也带来模型训练和运行速度都下降。尤其是当文本数据的数量急剧增长，单词的词表会变得巨大，这直接导致计算消耗十分大，包括计算复杂度太高所引起的运行耗时以及模型参数量太大所引起的内存占用过于庞大。为了解决这些问题，历史上人们开发了很多技术与模型，这些方法可以被分为三类：基于采样的近似算法，基于词表分解的算法，基于字母级别的编码模型。这些方法在一定程度上能缓解那些问题，但各自仍然存在一些缺陷。

本文首先概述语言模型的主要发展历史，详细阐述核心的建模方法——循环神经网络及其变种，解释大词表问题带来的重大挑战，列出了语言模型任务的国内外通用的标准数据集。然后上述所说的三大类前沿技术将会被总结分析优缺点，并重点讨论了基于词表分解的方法。最后，通过总结前人的方法，本文列出几个未来发展方向。

**关键词**：大词表问题，词表分解，循环神经语言模型， 自然语言处理

**Abstract**

Language model is a fundamental component in the field of neural language processing; its performance has direct influence on many tasks, e.g., machine translation and automatic speech recognition. With the propose of recurrent neural network and its inclusion in language model, language modeling by computer becomes more accuracy, yet producing the in-efficiency in the model’s training and operating. Especially when the text corpus growths explosively, vocabulary will become so huge that it causes both high computational complexity, which make it time-consuming to train as well as to infer, and excessive model size, which means the huge memory footprint. To tackle these problems, many techniques and models have been developed, mainly categorized as: sampling-based approximation, character level-based models, and vocabulary factorization-based methods.

In this paper we first review the development of the language model, containing the detail introduction of recurrent neural network and its variants, the explanation of the problem of over-large vocabulary, and some standard datasets. Afterwards, we explore those three kinds of advanced methods and analyze their merit and demerit. Finally, by summarizing the previous researches, we list some possible directions and the future works.

**Keyword:** over-large vocabulary, vocabulary factorization, RNN language model, natural language processing

**目 录**

1引言 1

1引言

近年来，互联网技术的快速发展，尤其是各大社交网络平台技术日趋完善，为人们营造了一个便于交流的世界。于是，在全球范围内互联网用户数量在不断增加，用户之间的交流、用户对互联网产品的评论以及各种信息的共享，这造成互联网上的数据爆炸式增长。其中，文本数据是用户信息传递的载体，如何从文本数据中自动挖掘有用的信息便成了一个各大互联网公司研究的热点。由此，对无标注的文本语言进行建模（Language Modeling, LM）就成了文本数据挖掘的一项非常重要的基础工作，很多文本挖掘应用场景，如文本检索【】，机器翻译【】和问答【】，都需要以语言模型为基本模块。因此，此研究课题是具有很强的实用价值和研究意义的。

在深度学习的概念提出之前，统计语言模型（Statistical Language Model, SLM） 是一种主要的对语言建模的方法，又称为N元语言模型（N-gram model）【】。自2006年Hinton提出深度学习（Deep Learning, DL）以来【】，将深度学习应用于自然语言处理（Natural Language Processing, NLP）任务中已经成为自然语言处理界的趋势，尤其是在语言模型中的应用。研究者提出可以用神经网络（Neural Network, NN）对语言进行建模，包括前馈神经网络模型（Feed-Forward Neural Network, FFNN）和循环神经网络模型（Recurrent Neural Network, RNN）【】。其中，RNN由于其特殊的网络结构能够将当前词的历史信息存储起来，并作为当前词预测的依据。该结构克服了N-gram语言模型无法利用距离当前词较远的词的信息。此外，在RNN对一个词序列进行建模时，离散的词被映射到连续稠密的词嵌入（Word Embedding）空间，在这低维空间进行计算学习词序列的特征来学习词与词之间的依赖关系。因此，循环神经网络语言模型（Recurrent Neural Network Language Model, RNNLM）在模型困惑度（Perplexity, PPL）和词识别率上都取得了最好的结果【】。

尽管