****

硕士学位论文开题报告及论文工作计划书

|  |  |
| --- | --- |
| 学 号： | 1701793 |
| 姓 名： | 张哲旸 |
| 导 师： | 朱靖波教授 |
| 学科类别： | 🗹工学 □全日制专业学位 |
| 学科/工程领域： | 计算机软件与理论 |
| 所属学院： | 计算机科学与工程学院 |
| 研究方向： | 机器翻译 |
|  |  |
| 拟选题目： | 面向神经机器翻译的后处理 |
|  | 方法研究 |
| 选题时间： | 2018 年 7 月 15 日 |

**东北大学研究生院**

**2018 年 9 月 1 日**

填表说明

1、本表应在导师指导下如实填写。

2、学生在开题答辩前一周，将该材料交到所在学院、研究所。

3、按有关规定，没有完成开题报告的学生不能申请论文答辩。

4、全文正文均用小四号宋体，单倍行距，段前段后间距为0，如果页数不够，可以整页扩页，其他格式要求参见《东北大学硕、博士学位论文格式》。

一、前期工作基础（本节可以整页扩页）

|  |
| --- |
| 课程学习及选题开题阶段，在导师指导下从事研究工作总结（不少于2000字）  本人在研一的学习过程学习该课题的相关课程，为该课题的开展打下了坚实的理论基础，并在导师的指导下阅读了相关资料以及从事了相关的研究工作。这部分内容我将从三个部分进行详细阐述，分别为相关课程的学习、文献阅读以及前期工作的研究。   1. **相关课程的学习** 2. 语言分析与机器翻译   该课程介绍了自然语言处理领域的任务以及相关理论知识，包括词性标注、分词、语言模型和机器翻译等概念，并且介绍了Niutrans统计机器翻译系统的使用方法。通过对统计机器翻译系统训练的实践，初步了解机器翻译的大致流程，为本课题的开展提供有效支持。   1. 最优化方法   神经网络的核心思想在于如何在一定条件下求得问题的最优解，神经网络本质上来说是最优化问题。最优化理论与方法作为一个重要的数学分支，它所研究的就是在众多的方案中怎么能找到最优、最好的方案。该课程最优化的基本理论和方法，介绍了线形规划、非线形规划和多目标规划的基本而常用的优化算法。介绍了梯度下降法、牛顿法、拟牛顿法、共轭梯度法等最优化方法，为学习神经网络，理解神经网络工作过程打好理论基础。   1. 人工神经网络理论知识   搭建神经机器翻译系统需要了解人工神经网络相关知识。包括前馈神经网络（Feedforward Neural Network, FNN），激活函数（activation function ），反向传播（Back propagation），梯度下降等相关知识，对人工神经网络理论知识的学习为本课题研究工作的展开，为如何搭建基线神经网络机器翻译提供了有效的途径。   1. 程序设计方法学   程序设计方法学是讨论程序的性质以及程序设计的理论和方法的一门学科， 是研究和构造程序的过程的学问。通过该课程的学习，为本课题中的编程需要奠定了良好的基础。是本课题能够顺利开展的一门必不可少的前修课程。   1. **相关文献的阅读**   在过去一年的研究生学习生活中，除了学习本研究课题相关基础知识外，还阅读了国内外相关参考文献，包括神经网络，统计机器翻译，神经机器翻译，机器翻译后处理等相关内容。   1. 神经机器翻译   机器翻译，又称为自动翻译，是利用计算机将一种自然语言(源语言)转换为另一种自然语言(目标语言)的过程。在机器翻译发展期间先后提出了各种机器翻译技术，包括：基于规则的机器翻译、基于统计的机器翻译、基于神经网络的机器翻译。神经机器翻译（Neural Machine Translation ,NMT）是以神经网络作为基础的机器翻译，已经全面超越此前以统计模型为基础的统计机器翻译（Statistical Machine Translation, SMT），并快速成为在线翻译系统的主流标配。相比于传统的统计机器翻译（SMT）而言，神经机器翻译能够训练从一个序列映射到另一个序列的神经网络，输出的可以是一个变长的序列。这在翻译、对话和文字概括方面能够获得非常好的表现。  神经机器翻译使用“编码-解码”（Encoder-Decoder）框架。对于一个源语言句子，首先提取源语言中信息，将它映射为一个连续、稠密的向量，这一过程是通过编码器Encoder实现的，然后再将这个向量转化为目标语言的句子，这一过程通过解码器Decoder实现，从而完成对语言的翻译。如图1所示，X1、X2、X3、X4是源语句子X中的每一个单词，c为将输入句子X通过非线性变换转化为的中间语义表示，解码器Decoder根据源语句子X的中间语义表示c和之前已经生成的历史信息Y1,Y2…Yi-1来生成i时刻要生成的单词Yi。  C:\Users\Dante_Lab\AppData\Local\Temp\1535341587454.png  图1 Encoder-Decoder 框架  机器翻译需要模型能够有效对变长序列进行处理。在神经网络中，对于变长序列的问题最经典的解决方案之一就是通过循环神经网络对问题进行建模。由于循环神经网络（Recurrent Neural Network, RNN）在神经网络结构中引入了时序的概念，因此对于变长序列的处理有着得天独厚的优势。 基于RNN的神经机器翻译模型如图2所示。  C:\Users\Dante_Lab\AppData\Local\Temp\1535354898101.png  图2 RNN-based NMT  虽然循环神经网络在机器翻译上已取得很好的效果，不过由于循环神经网络结构的原因还存在一些问题，比如RNN对层次信息的表示能力和卷积能力都存在一些不足 。因此产生了一些如Deep RNN、Bidirectional RNN和Hierarchical RNN的变种，都希望通过更层次化的网络结构来弥补这种不足。此外，RNN的并行度很低。它的当前时刻的隐藏层的状态依赖于之前时间的状态，并且训练和解码的并行度都很低。 为了解决上面的问题，Google提出了一种新的神经网络的架构——Transformer，将自注意力机制应用到神经机器翻译中，完全消除了模型对RNN和CNN等单元的依赖。由于自注意力机制本身的特点，模型能够有效处理一句话中距离较远的两个关联单词之间的关系，不同位置的信息传递距离仅为O(1)，这样可以非常充分地表示序列中不同位置之间的复杂关系，使得该模型对于长句子具有更好的翻译效果。Transformer模型结构如图3所示。    图3 Transformer-based NMT   1. **前期工作总结**   在过去的一年里，本人一直根据导师的指导，从事机器翻译，神经网络等相关研究工作。   1. 在实验室参加的WMT 2018国际翻译大赛和CWMT 2018 全国机器翻译研讨会中，本人负责了机器翻译的后处理部分工作，具体工作分为标点符号后处理以及命名实体错误翻译后处理。   1）符号后处理  在我们的日常表示中，英文的标点符号是形如“,.?”这样的半角标点符号，而中文一般使用形如“，。？”这样的全角标点符号，在英中机器翻译系统中，经常会出现经过机器翻译系统翻译后的译文中大量使用英文中的半角标点符号，影响读者阅读体验并且导致机器翻译自动评估标准——BLEU (Bilingual Evaluation Understudy)的分数过低。在这里本人通过使用符号替换，正则表达式等方式将中文翻译结果里面的半角标点符号替换为全角标点符号。在英中的BLEU分数上有着显著的提高。   * 1. 命名实体错误翻译后处理   在机器翻译过程中，一些原文中可能会有目标语言的命名实体，这些命名实体在解码之后可能会翻译错误，比如将“Wirecard”被翻译成“Wiretard”,针对这一情况，提出一种基于原文的后处理算法，如图4所示，使用原文中的命名实体对译文中错误翻译的命名实体进行替换。  C:\Users\Dante_Lab\AppData\Local\Temp\1535356489888.png  图4 后处理算法  在实际应用中，我们使用如下公式计算归一化编辑距离：  （1）  在计算相似度的时候，我们以编辑距离为基础，并且考虑了长度的影响，D(x,y)代表x、y之间的编辑距离，Lx是x的长度。我们使用源语句中的专有名词替换译文中相似度最高的单词。  由于包含目标语言的命名实体的句子的数量较少，这个方法在BLEU上没有显著的改进，但是我们发现它在人工评估是十分有效的，特别是当原文中包含较多人名以及品牌名的时候。   1. 数字翻译问题后处理   在对机器翻译的系统的翻译结果观察对比中，本人发现机器翻译对简单数字翻译效果很好，不过在对复杂数字翻译时经常会产生如下例所示的错误翻译现象，包括数字翻译错误以及单位翻译错误两种情况。  包括人员支出3,706.87万元。  includes $ 37.067 million for personnel.  产生这一现象的原因包括复杂数字的训练语料中并不多见、分词错误、BPE（Byte Pairs Encoding）切分错误等可能性。数字翻译错误问题十分影响机器翻译系统使用者的用户体验。为提高机器翻译系统对复杂数字的翻译质量，本人提出一个通过简单数字替换方式处理数字翻译问题的后处理方法。  1.复杂数字替换为简单数字  2.翻译替换后的新句子  3.通过attention对齐信息还原  4.数字正规化  首先找出待翻译语句中数字部分，确定哪些数字是机器翻译系统不易翻译的复杂数字。然后将一个将长的、复杂的数字替换为较短的、简单的易翻译数字。比如将“省 统计局 2015 年 收入 预算 14,713.35万 元”这句话中的“14,713.35”替换为“100”。将替换后的新语句“省 统计局 2015 年 收入 预算 100万 元”送入机器翻译系统进行翻译。在翻译过程中，输出原文中各个单词对应的译文中单词的attention对齐概率，如图5所示，最后通过attention对齐信息将翻译后的简单数字通过计算还原为复杂数字，最终得到不存在数字翻译问题的译文。  C:\Users\Dante_Lab\AppData\Local\Temp\1535356172564.png  图5 “100”万对应的attention对齐概率 |

二、选题依据（本节可以整页扩页）

|  |
| --- |
| 课题背景、选题依据、课题研究目的、理论意义和应用价值（工学硕士）/工程背景和实用价值（专业学位硕士）（不少于1000字）   1. **课题背景**   随着全球化的深入发展，世界各国的政治、经济、文化交流与合作不断深化，跨国人口流动日益频繁。来自世界各地的人交流日益增多，然而语言仍然是全球交流的主要障碍。传统的解决方法是采用人工翻译，但是人工翻译佳哥昂贵、翻译速度慢，不能再根本上解决语言障碍问题。  随着计算机科学技术的发展，机器翻译作为自然语言处理研究的重要组成部分越发受到人们关注。经过了几十年的努力，机器翻译工作取得了巨大的进展，机器翻译可以实现世界上不同国家不同语言间的低成本交流，其主要优点体现为：   1. 成本低。相对于人工翻译来说，机器翻译的成本要低很多。机器翻译需要人工参与的程序其实很少，基本上由计算机自动完成翻译，大大降低了翻译成本。   2）易把控。机器翻译的流程简单快捷，在翻译时间的把控上也能进行较为精准的估算。  3）速度快。计算机程序的运行速度非常快，其速度是人工翻译速度不可比拟的。  由于这些优点，机器翻译在这几十年来得到了快速的发展。此外，网络化和国际化对翻译的需求日益增大、人工智能与深度神经网络的快速发展、现如今日益增长的互联网规模能有效地将网络上的数字资源加以利用，能够很好地为自然语言处理以及机器翻译提供强有力的数据支撑等都促进了机器翻译的蓬勃发展。  早期的机器翻译是基于经验主义的翻译方法，主要是基于实例和基于统计的方法，特点是注重大规模语料库的建设，开始了针对大规模的真实文本处理。同时，这一阶段的研究工作开始解决一个比文本翻译更加复杂和艰难的问题——语音翻译。而且由于Internet上的机器翻译系统具有巨大的潜在市场和商业利益，此时网上翻译机器系统也进入了实用领域的新突破阶段。  机器翻译功能越来越强大，从最初只能进行简单的单词翻译，到之后可以翻译出基本符合语法的句子，慢慢可以翻译具有一定逻辑性的句子，现在部分软件已经可以自主联系上下文进行翻译，翻译结果的准确性与可读性都已经取得了非常大的进步。  近年来，加入了“深度学习技术”等人工智能的神经机器翻译（Neural Machine Translation ，NMT）已经不止于简单的将一个个单词翻译成另一种语言，而是可以像人工翻译一样，不断向前回顾理解结构复杂的句子，同时联系上下文进行翻译。  基于人工神经网络的机器翻译（ Neural Machine Translation ）技术核心是一个拥有海量结点（神经元）的深度神经网络，可以自动的从语料库中学习翻译知识。一种语言的句子被向量化之后，在网络中层层传递，转化为计算机可以“理解”的表示形式，再经过多层复杂的传导运算，生成另一种语言的译文。实现了“理解语言，生成译文”的翻译方式。这种翻译方法最大的优势在于译文流畅，更加符合语法规范，容易理解。相比之前的翻译技术，质量有“跃进式”的提升。和其他深度学习领域一样，神经机器翻译系统的翻译性能受到语料数据影响，并且每次训练时间长，训练效率低、修改困难等问题，越来越多的研究人员在尝试提高各种机器翻译性能的方法来辅助神经机器翻译。  机器翻译后处理作为一个对人工评价有着明显提高的方法，在WMT、CWMT等国内外顶级机器翻译评测中被大量应用，然而相关的理论研究、论文等文献数量十分稀少，而且现有的研究都是基于规则的后处理，面向的机器翻译系统是传统的统计机器翻译系统。现有的研究主要是对统计机器翻译里面的unknown words(UNK)问题，即带翻译的原文中出现了训练语料中从未见过的单词。传统的后处理方法处理诸如数字翻译等这些训练语料无法全部包含的词汇时，会将待翻译句子中的数字部分使用一个特殊的标签代替，然后这个标签在经过统计机器翻译系统时候不会被翻译，在得到带数字标签的译文后，后处理技术使用事先编写好的数字翻译模块对原文中数字进行翻译，对译文的数字标签替换，得到完整的译文。  除了针对统计机器翻译里面的数字翻译问题以外，也有部分研究者针对一些翻译性能比较差的统计机器翻译系统使用后处理方法修正某些单词。有研究人员使用后处理方法处理汉蒙机器翻译系统的动词词形错误问题。这些后处理操作十分依赖人工的检验，通过翻译人员的知识对这种性能较差的翻译系统的译文进行人工检查，发现各条译文中普遍存在的错误问题，然后对这些错误问题编写规则，进行修改。  后处理方法已经在在WMT、CWMT等国内外顶级机器翻译评测中得到大量应用，后处理方法的有效性已被证明，然而后处理方法相关的理论研究、论文等文献数量较少，使用的手段也过于局限，值得研究人员深入研究。   1. **选题依据**   由现在，机器翻译已经成为世界自然语言处理研究的热门，基于人工神经网络的机器翻译（ Neural Machine Translation ）在国际机器翻译公开评测中性能已被证明可以达到或者超出传统统计机器翻译方法。基于人工神经网络的机器翻译能够有效地翻译语法复杂的语言并且可以考虑完整的句子，而不仅仅是一串单词。 不过对比传统统计机器翻译方法，神经机器翻译虽然在翻译性能上更出色，但其译文仍存在如下几个问题：词语表规模受限问题、源语言翻译覆盖问题 、如何保证对原文的忠实度问题以及如何引入先验知识和外部知识提高翻译性能 。  机器翻译的流程可以分为预处理、核心翻译、后处理三块，当前大部分研究人员主要着眼于提升核心翻译性能，忽略了后处理部分对机器翻译性能的提高。然而核心翻译部分的提高并不能很好的使用先验知识来提高翻译性能，当前性能最好的机器翻译系统的核心翻译部分都是使用神经网络，然而随着网络规模的扩大，网络中的参数也会越变越多，网络的计算速度变得越来越慢，针对部分内容而专门优化神经网络费时长久、开销巨大且很难产生显著效果，使用后处理技术提高翻译性能则不会有此类问题。传统的后处理模块是将翻译结果进行大小写的转化、建模单元进行拼接，特殊符号进行处理，使得翻译结果更加符合人们的阅读习惯。除此以外，机器翻译后处理模块功能有着很大的扩展空间，比如可以使用原文-译文词汇对齐信息、针对命名实体进行优化等。机器翻译后处理技术可以使用很小的时间开销代价完成对机器翻译译文的优化，并且可以使用先验知识和外部知识提高翻译性能，有着很好的研究前景。   1. **课题研究目的**   本课题的研究目的就是解决神经机器翻译存在的源语言翻译覆盖不全问题、无法引入先验知识和外部知识提高翻译性能以及修改模型后在训练时间开销过大等问题，使其能够在不增加神经网络训练时间的情况下很好针对特定方面提高机器翻译性能，增加译文的可读性，并且可以保证译文质量不会因为后处理操作有所下降。   1. **理论意义及应用价值**   目前对神经机器翻译后处理操作的研究工作正处于起步阶段，虽然部分研究人员提出针对汉蒙翻译等语种的机器翻译后处理操作，不过这些后处理操作打多是基于规则的、针对某些语种的，有着针对性过强、适应性不足的缺点，并且并不能解决神经机器翻译无法使用外部知识这一缺点，对神经机器翻译后处理深入研究可以提高机器翻译性能，使用外部知识提高机器翻译效果，对神经机器翻译的发展产生深远的影响。 |

三、文献综述（本节可以整页扩页）

|  |
| --- |
| 国内外研究现状、发展动态描述（不少于1000字）；所阅文献的查阅范围及手段，附参考文献（不少于10篇，其中近3年文献不少于5篇，英文文献不少于3篇，全部按照标准格式列出，并在文中顺序标注）   1. **国内外研究现状**   机器翻译研究如何利用计算机自动地实现不同语言之间的相互转换，是自然语言处理和人工智能重要研究领域，也是目前互联网常用服务之一。如Google翻译、百度翻译、微软Bing翻译等，都提供多种语言之间的在线翻译服务。虽然机器翻译译文质量与专业译员相比仍有较大差距，但是在一些对译文质量要求不太高的场景下，或者是在特定领域翻译任务上，机器翻译在翻译速度上具有明显优势，仍然得到广泛应用。  机器翻译的研究始于20世纪50年代，是人工智能最早的研究领域之一。当时的主要做法是依靠人工编写翻译规则，将源语言的句式和词汇按照、固定规则转换为目标语言。然而人们很快认识到语言的复杂程度是难以用规则涵盖的。到20世纪80年代为止，机器翻译只在天气预报等语法简单、词汇固定的个别领域实现了应用。20世纪90年代起，IBM提出了统计机器翻译[1][2]，他们提出的IBM模型实现了在平行预料上自动学习单词之间的对应，再与语言模型结合，实现了基于单词的机器翻译系统。在此基础上，基于短语的机器翻译[3][4]将单词之间的对应扩展到词组之间的对应，实现了当时最优的翻译效果，随即被工业界广泛采纳。 近几年，随着深度学习的崛起，基于神经网络的机器翻译方法获得迅速发展[5][6]，与传统的统计翻译模型相比，2014年提出的循环神经网络Seq2Seq[7][8]在概念上要简单得多。在统计翻译模型中，模型的训练步骤可以分为预处理、词对齐、短语对齐、抽取短语特征、训练语言模型、学习特征权重等诸多步骤。Seq2Seq模型的基本思想非常简单一一使用一个循环神经网络读取输入句子，将整个句子的信息压缩到一个固定维度的编码中；再使用另一个循环神经网络读取这个编码，将其“解压”为目标语言的一个句子。这两个循环神经网络分别称为编码器（Encoder）和解码器（Decoder），这个结构也称为Encoder-Decoder结构[7]。  2017年，谷歌提出了一新框架Transformer[9]，完全基于注意力机制，实验表明，这一模型在机器翻译任务上表现极好，具有可并行化的特点，并且大大减少训练时间，而且模型性能相比较RNN有较为明显的提升。Transformer也基于Encoder-Decoder结构，不过与循环神经网络不同的是Transformer采用了一种全新的注意力机制，包含Encoder端、Decoder端的自注意力机制以及Encoder-Decoder的联合注意力机制。Encoder部分将输入序列表示成一个带有语义的向量，Decoder部分是以encoder生成的向量作为输入“解码”出目标文本序列。Transformer模型由于没有使用循环和卷积结构，在使用序列的顺序信息时需要将相对位置信息以及绝对位置信息注入到模型中。  机器翻译后处理工作最早可以追溯到20世纪50年代，也就是在的机器翻译研究工作的早期，研究人员主要依靠人工编写的规则等外部知识来编写机器翻译系统，这部分使用外部知识提高翻译性能的工作和机器翻译后处理操作的部分想法是一致的，在统计机器翻译系统和神经机器翻译被大幅推广之后，关于机器翻译后处理研究有所减少，目前，有研究人员使用后处理操作处理汉蒙翻译的词性错误等问题[11][12][13][14][15]，以及解决词表外单词（Out-of-Vovabulary,OOV）翻译问题[16][17][18][19][20]，也有部分研究人员在开发开源的针对某一语种的后处理方法系统[21]。除此之外，在WMT2017国际翻译大赛上，多个参赛队伍使用后处理提高译文质量[22][23][24][25]。   1. **所阅文献的查阅范围及手段**   本课题所查阅的文献主要来自于人工智能、自然语言处理等领域会议论文集，NIPS、IJCAI、ACL、EMNLP、WMT等。前两二者主要围绕机器学习和人工智能相关领域进行探讨，后三个则主要围绕自然语言处理相关任务进行讨论，其中每个会议发表的论文都是针对近几十年来相关方向上最前沿、最值得探讨的问题。这些论文为本课题所实现的神经机器翻译后处理提供支撑。文献主要通过谷歌学术搜索、每篇论文后的参考文献指引、中国知网论文检索和老师推荐获得。   1. **参考文献**   [1] Brown P F, Cocke J, Pietra S A D, et al. A statistical approach to machine translation[J]. Computational Linguistics, 1990, 16(2):79-85.  [2] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2):263-311.  [3] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003:48-54.  [4] Chiang D. A Hierarchical Phrase-Based Model for Statistical Machine Translation.[C]// ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa. DBLP, 2005:263--270.  [5] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. 2016.  [6] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436.  [7] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. 2014, 4:3104-3112.  [8] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.  [9] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[J]. neural information processing systems, 2017: 5998-6008.  [10] Liu N F, May J, Pust M, et al. Augmenting Statistical Machine Translation with Subword Translation of Out-of-Vocabulary Words[J]. arXiv preprint arXiv:1808.05700, 2018.  [11] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980.  [12] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Open- NMT: Open-Source Toolkit for Neural Machine Translation. In Proc. of ACL.  [13] Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In Proc. of EACL, pages 187–193.  [14] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In Proc. of the First Workshop on Neural Machine Translation, pages 28–39.  [15] Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates Can Improve Statistical Translation Models. In Proc. of NAACL, pages 46–48.  [16] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 10:707.  [17] Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In Proc. of NAACL, pages 104– 111.  [18] Minh-Thang Luong and Christopher D. Manning. 2016. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In Proc. of ACL, pages 1054–1063.  [19] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attentionbased Neural Machine Translation. In Proc. of EMNLP, pages 1412–1421.  [20] Pranava Swaroop Madhyastha and Cristina Espa˜na Bonet. 2017. Learning Bilingual Projections of Embeddings for Vocabulary Expansion in Machine Translation. In Proc. of the 2nd Workshop on Representation Learning for NLP, pages 139–145.  [21] Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In Proc. of EMNLP, pages 381–390.  Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. ArXiv:1309.4168.  [22] Huck M, Riess S, Fraser A. Target-side word segmentation strategies for neural machine translation[C]//Proceedings of the Second Conference on Machine Translation. 2017: 56-67.  [23] Nadejde M, Reddy S, Sennrich R, et al. Predicting target language CCG supertags improves neural machine translation[J]. arXiv preprint arXiv:1702.01147, 2017.  [24] Chatterjee R, Negri M, Turchi M, et al. Guiding neural machine translation decoding with external knowledge[C]//Proceedings of the Second Conference on Machine Translation. 2017: 157-168.  [25] Ding S, Duh K, Khayrallah H, et al. The JHU machine translation systems for WMT 2016[C]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016, 2: 272-280. |

四、研究内容（本节可以整页扩页）

|  |
| --- |
| 1．研究构想与思路、主要研究内容及拟解决的关键问题（不少于1000字）  现有的机器翻译后处理技术的研究思路主要是考虑一些特定语言的语法规则，比如对汉蒙机器翻译系统的蒙语动词部分进行后处理操作。一方面，主要处理在机器翻译语料中未出现的词，如语料中无法全部包含的数词，现有的机器翻译后处理技术的研究思路主要使用基于规则的方法实现数词的自动翻译。另一方面，主要处理一些位于句子中特殊位置的词的词形错误情况，针对这一问题，现有研究思路采用基于规则的手段，使用词干提取等方法辅助对错误单词词形进行后处理，修改成正确的形式。  这些研究方法错在一些不足之处，首先是针对性过强，适应性太弱问题。传统的基于规则的后处理技术只是针对某一种语种的语法规则简单的进行基于规则的后处理操作，在其他语种上无法使用。而且处理效果十分依赖于词干提取等相关方法。  其次是过于生硬的将一些数词翻译替换为基于规则的翻译，无法充分利用机器翻译系统的性能，在单位处理容易出现错误翻译问题，以及银行卡号、电话号等非数词的数字串翻译上容易出现错误翻译问题。  最后，之前的研究方法后处理操作主要是基于规则层面，不能充分的利用机器翻译系统性能，针对单一语言，拓宽到多语言上代价很大，而且无法使用外部知识和神经网络内部知识。需要一种研究思路来使后处理操作可以覆盖多语种，使用更多信息进行有效的译文修正操作。   1. **研究构想与思路**   本课题将采用谷歌推出的TensorFlow人工智能学习系统进行实现，搭建完全基于 attention的Transformer网络架构机器翻译系统，在机器翻译领域，Transformer性能已被证明优于循环与卷积神经网络。并且可以输出对齐信息等神经网络内部信息供后处理操作使用。  我们采取基于Transformer网络架构神经机器翻译系统作为我们的研究的基线系统，并在基线系统的相同条件设置下开展相关实验。为了评估经过后处理操作后的译文的优劣，还需要设计一个评估策略，包括后处理所用时间、修改幅度以及BLEU值等。基于以上构想，本课题的具体研究思路如下：  1）确定合适的神经机器翻译系统，作为本课题研究的基线系统，相关实验在基线同等条件下展开。  2）构建合适的测试数据集，尽量包含较多机器翻译中情况中会出现的双语句对情况，并设计一个合理评估策略。在测试数据集的选择上，部分采用mt06、mt08等开源测试集内容，可以和前人工作对比，部分采用互联网上的真实数据，代表性较强。  3）在获得基线系统和合理测试数据集的基础上，首先使用原文内信息对机器翻译系统的译文进行后处理操作。原文内信息是各个语种之前都拥有的通用信息，使用原文内信息的后处理方法，不受语种的限制，可以很方便的拓展到其他语种之中。使用原文的后处理方法可以处理诸如词形错误、专有名词错误、部分数字翻译问题等。  4）使用先验知识和外部知识对译文进行后处理操作。本研究课题准备使用句法树的外部知识应用到神经机器翻译后处理之中，句法树可以分析句子成分，可以解决部分语法错误，除此以外，句法树对词性的标注应用到后处理之中可以处理诸如时态错误、动副词词形错误等语法错误问题。  5）使用神经机器翻译模型输出的对齐关系等相关信息进行后处理操作。Attention机制是主流神经机器翻译系统都会使用的关键机制，使用attention对齐规律可以获得原文和译文中各个单词的逐一对照概率，可以分析出各个单词的对应关系，使用这些知识可以让后处理能够处理数字翻译问题、翻译丢失问题，此外，本课题研究还准备在后处理方法中提高attention对齐概率的准确性，然后将提高后的attention概率应用到Decoder等神经机器翻译内部结构中，提高神经机器翻译整体性能。   1. **主要研究内容及拟解决的关键问题**   本本课题的主要研究内容以及拟解决的关键问题有以下几个方面：   1. 神经机器翻译基线系统的搭建   搭建神经机器翻译系统是本课题研究的基础。使用TensorFlow实现完全基于attention 的Transformer网络架构，它仍然是当前效果最好的神经机器翻译基本架构，在这个基础上做神经机器翻译的后处理操作更具有代表性。   1. 测试数据集的构建   构建合理的测试数据集是本课题研究的基础。现有的开源测试数据集大多为了评判机器翻译的整体性能，对后处理操作的针对性不强，这正是在当前研究该方向时比较欠缺的地方。合理的测试数据集可以反应后处理工作的效果，包含较多情况的数据集也可以为后处理工作提供研究方向。   1. 合理的后处理操作评估策略   当前并没有一个完善的后处理苹果策略，后处理操作大部分是对部分译文进行操作，且操作幅度为数个单词，一个合理的评估策略可以反应后处理操作的力度和效果。   1. 研究合理的后处理方法，构建后处理系统   研究后处理操作的有效手段是本课题研究的关键。很多机器翻译系统并没用使用后处理系统，本课题研究将对编辑距离、向量距离等后处理中使用的方法进行实验对比，并且使用神经机器翻译模型输出的对齐关系等相关信息进行后处理操作、使用先验知识和外部知识对译文进行后处理操作。开发具有实际使用意义的、适应绝大部分语种后处理系统是本课题的研究内容关键。 |
| 2．拟采取的研究方法、技术路线、实施方案及可行性分析   1. **拟采取的研究方法**   使用原文信息的后处理方法：机器翻译是将输入的源语翻译为目标语，主流机器翻译系统使用的Encoder--Decoder框架的工作流程是编码器Encoder对输入的源语句子进行编码，将输入的源语句子通过非线性变换转化为中间语义表示，之后解码器Decoder根据源语句子中间语义表示来生成目标语译文，看起来就是整个机器翻译系统根据输入的源语句子生成了目标句子。在这过程中，所有的源语信息都会进行编码生成中间语义信息，而在实际语言使用中，一些源语里面的信息是不需要进行翻译的，比如一些类似于“iPhone”、“Windows 10”、“Niutrans”的品牌名、系统名，而且很多机器翻译系统的训练数据无法包含所有的品牌名，源语句子在经过BPE等单词切分方法后，在这些品牌翻译时很有可能产生错误，后处理操作将对这些情况进行处理，优化译文。在处理时，寻找对应的翻译错误的单词的方法不同，对后处理效果会有影响，具体方法的选择将在后续研究中进行探讨。  使用神经网络信息的后处理方法：一般来说，神经机器翻译系统的目标是将输入的源语句子翻译为目标语，神经网络中产生的各种信息都是为其服务的，然而有时译文的质量不佳可是使用后处理操作利用神经网络中的信息来提升译文质量。比如使用神经网络中的对齐信息确定输入的原文句子中的单词对应译文中的哪个单词等。采取何种信息、如何在不影响神经网络运算速度的情况下在网络中提取信息应用到后处理操作中将在后续研究中进行探讨。  除上述两个研究重点外，本课题还将从理论的角度对后处理操作可以进行的处理的情况进行分析，并尝试使用外部知识、先验知识应用到后处理操作中提高翻译性能，在对后处理操作的研究中深入考虑是什么问题导致译文翻译质量不佳，如何在神经网络中进行调整避免这一现象的发生。   1. **可行性分析**   我认为本课题的研究方案以及技术路线在选题、前期研究基础以及实验条件等方面具有可行性，具体内容包括：  选题：本课题研究内容贴合实际需求，神经机器翻译在得到广泛应用的同时，如何提高训练数据中很少或者未出现过的原文的翻译质量一直是一个让人头疼的问题，如何对长数字、专利号等长度过长的单词进行正确的翻译也一直困扰机器翻译研究人员的问题，而且神经网络在具有优秀性能的同时，也让机器翻译很难使用先验知识、外部知识等内容提高翻译质量，这些都是神经机器翻译目前极其重要的问题。因此针对神经机器翻译系统后处理的相关研究十分重要，可以有效解决困扰机器翻译研究人员的数个翻译问题，并且不会导致翻译时间过长等情况出现。  前期研究基础：在过去一年的研究过程中，本人所学习的课程（如语言分析与机器翻译、最优化方法学等）可为本课题的开展提供理论支撑，同时在课下的时候阅读了关于机器翻译以及神经网络等领域相关文献资料，对课题有了一定的了解。此外参加了WMT 2018、CWMT 2018机器翻译大赛，对机器翻译以及后处理的实际应用有所了解，并参与完成了三篇机器翻译和自然语言处理领域的论文。  实验条件：本课题中涉及的实验项目依托东北大学自然语言处理实验室，能够使用其提供的多GPU设备进行试验。同时实验室提供本课题所需要的训练数据和性能优秀的机器翻译系统，可以验证后处理操作在实际应用是否有效。 | |

五、预期研究成果（本节可以整页扩页）

|  |
| --- |
| 对所研究的成果进行阐述，同时要对与前文研究内容的相关性及与前人（他人）研究成果的差异性进行描述   1. **预期研究成果**   本课题预计在结束后将产生如下输出：   1. 可在实际中使用的机器翻译后处理模块   本课题针对机器翻译译文进行后处理操作，在课题结束时，预计可产出一套可实际应用的机器翻译后处理模块，其功能主要支持数字翻译问题后处理、专有名词后处理等，独立性高，可连接到不同机器翻译系统之后，并且可应用到多个语种之中。   1. 后处理效果分析方法   提出一套行之有效的理论方法，能够针对判断机器翻译后处理操作是否产生正向效果。   1. 神经机器翻译后处理方案   提出一套效果良好的神经机器翻译后处理方案，可以对不同神经机器翻译系统的译文进行后处理操作，并且不受语种限制。   1. 能够在相关会议期刊上发表高水平论文   能够在机器翻译、神经网络以及自然语言处理等相关领域发表论文，至少1篇以上。   1. **与前人研究成果的差异性**   本课题和前人工作差异如下：   1. 与前人工作面向的任务不同   本课题旨在探究神经机器翻译的后处理方法，前人的研究思路主要是基于统计方法的机器翻译系统的后处理技术。只有部分比赛中使用神经机器翻译系统时提到了后处理技术。无论是何种后处理方法，应用于神经机器翻译的相关文献并不多见。神经机器翻译系统整体性能由于统计机器翻译系统，在神经机器翻译系统上研究后处理操作方法更有实际价值。   1. 与前人工作采取的手段不同   本课题旨在探究神经机器翻译的后处理方法，前人的研究思路主要是考虑一些特定语言的语法规则使用规则的手段进行后处理操作。一部分研究工作旨在处理在机器翻译语料中未出现的词，如语料中无法全部包含的数词，现有的机器翻译后处理技术的研究思路主要使用基于规则的方法实现数词的自动翻译。也有部分研究人员主要处理一些位于句子中特殊位置的词的词形错误情况，采用基于规则的手段，使用词干提取等方法辅助对错误单词词形进行后处理。而使用其他信息进行后处理操作的方法，在神经机器翻译的相关文献并不多见。本课题的研究思路是使用原文信息、神经网络内的信息等有价值的信息来进行后处理操作，预期效果更好。   1. 与前人工作针对范围不同   由于不同语言语法规则的不同，前人研究思路使用规则的方法进行后处理操作时，不同语种要写不同的规则来进行后处理操作，前人编写的后处理模块适应范围很窄，本课题研究思路不局限于基于规则的后处理操作，使用各个神经机器翻译系统都具有的原文信息、神经网络内的信息进行后处理操作，使用范围不局限于单一语言，适应性更好。 |

六、研究条件（本节不允许扩页）

|  |
| --- |
| 1．所需实验手段、研究条件和实验条件  本课题所需实验条件主要包括硬件和软件两个部分。其中硬件需要多台载有多GPU卡的服务器，这部分东北大学自然语言处理实验室提供，共计4台符合条件的设备进行实验（两台4卡GTX TITAN，两台6卡TITAN X，显存均为12GB）。软件部分需要使用TensorFlow深度学习框架，搭建神经机器翻译系统针对其中的后处理部分进行开发和实验。  2．所需经费，包含经费来源、开支预算（工程设备、材料须填写名称、规格、数量）  无 |

七、工作计划（本节不允许扩页）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 阶段及内容 | 工作量估计  （时数） | | 起止日期 | 阶段研究成果 |
| 1 | 课题方向研究和选择 | 480 | | 2018.5.1-2018.7.31 | 收集资料，确定研究方向 |
| 2 | 撰写开题报告，明确研究内容 | 160 | | 2018.8.1-2018.8.31 | 撰写开题报告，完成开题答辩 |
| 3 | 基于Transformer搭建  神经机器翻译基线系统 | 480 | | 2018.9.1-2018.11.30 | 神经机器翻译系统 |
| 4 | 从理论的角度分析后处理能处理情况、不同方法对后处理效果的影响 | 320 | | 2018.12.1-2019.1.31 | 提出一份切实可用的后处理方案 |
| 5 | 使用源语中信息、神经网络网络中信息进行神经机器翻译后处理实验，分析效果对存在的问题提出解决方案 | 320 | | 2019.2.1-2019.3.31 | 提出一份切实可用的后处理模块 |
| 6 | 使用外部知识进行后处理实验，并分析造成译文错误的原因，提出解决方案 | 640 | | 2019.4.1-2019.7.31 | 使用外部知识的后处理模块、神经机器翻译优化方案 |
| 7 | 将上述解决方案下的神经机器翻译  后处理研究成果进行总结，投稿一篇论文 | 320 | | 2019.8.1-2019.9.30 | 神经机器翻译后处理实验结果总结，论文一篇 |
| 8 | 总结当前实验结果的不足之处并完成神经机器翻译系统后处理模块的优化工作 | 100 | | 2019.10.1-2019.10.14 | 神经机器翻译系统后处理模块的优化 |
| 9 | 对研究工作进行整理并撰写论文 | 480 | | 2019.10.15-2019.12.31 | 完成毕业论文 |
| 合计 | 3300 |