中图分类号：TP183

论 文 编 号 ：10006 ZY1624108

北京航空航天大學

硕 士 学 位 论 文

基于神经网络的
中文自动分词方法研究

作 者 姓 名　　魏 璋 兴

学 科 专 业　　工 业 工 程

指 导 教 师　　于 雷　副 教 授

培 养 学 院　　中 法 工 程 师 学 院

Research of Chinese Word Segmentation Method Based on Neural Architecture

A Dissertation Submitted for the Degree of Master

Candidate： Zhangxing Wei

Supervisor： Prof. Lei Yu

Sino-French Engineer School

Beihang University, Beijing, China

中图分类号：TP183

论 文 编 号 ：10006 ZY1624108

硕 士 学 位 论 文

基于神经网络的中文自动分词方法研究

作者姓名	魏璋兴	申请学位级别	全日制工程硕士
指导教师姓名	于雷	职　　称	副教授
学科专业	计算机技术	研究方向	自然语言处理
学习时间自	2016 年 9 月 15 日	起至	2019 年 1 月 15 日止
论文提交日期	2018 年 11 月 12 日	论文答辩日期	2018 年 12 月 7 日
学位授予单位	北京航空航天大学	学位授予日期	年　　月　　日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：_____ 日期：　　年　　月　　日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名：_____ 日期：　　年　　月　　日

指导教师签名：_____ 日期：　　年　　月　　日

摘要

随着近年来自然语言处理（Natural Language Processing）和机器学习（Machine Learning）技术的不断发展，大量的应用需要对利用计算机对文字进行自动处理、理解和研究。而词被认为是汉语中最小且具有语义的语言单位。在此背景下，自动分词（Chinese Word Segmentation）作为中文自然语言处理的基本工作就显得格外重要。

本文从国内外中文分词研究背景、最新成果以及遇到的问题出发，对文字的处理方式和向量表示技术进行了深入地探究。除了基于结合词典和规则匹配的传统分词方法，着重对基于字标注的统计学习方法和基于词的神经网络搜索方法进行了分析和比较。

通过对现有的多种分词方法的对比研究，本文总结了各种方法的优缺点，发掘研究趋势。在近年来使用循环神经网络和卷积神经网络对字序列建模的分词方法迅速发展的基础上，创新性地提出使用纯注意力机制结合语言模型进行自动分词的方案，一定程度上解决了序列中长距离语义依赖的难题，并取得了泛化性能和分词速度上的提升。在多个公开数据集进行实验，获得了与先进分词方法十分接近的测试成绩，进一步指出了深度学习在中文分词领域可能具有的发展方向和潜力。

关键词：中文分词，机器学习，语言模型，神经网络，注意力机制

Abstract

The increasingly development of Natural Language Processing and Machine Learning technologies has led to a huge demand of automatic comprehension and generation of natural language by the machine in recent years, meanwhile, word is recognized as the smallest but meaningful unit in Chinese language.

Thus the Chinese Word Segmentation, seen as a preliminary step for most NLP tasks, has become more and more significant. In this paper, we deeply survey the different ways of extracting features of words to vector representation as well as the most frequently used CWS methods. In addition to those traditional methods relied on multimode matching using dictionary and rules, we focus especially on those character-based tagging methods and word-based deep learning methods.

By analyzing and comparing these different methods, we discuss the pros and cons of each of them. In the light of the current trend in NLP, we creatively proposed a method based entirely on attention mechanisms, combining with language model, dispensing with complex recurrent or convolutional neural networks which are dominant sequence neural models for CWS. With the help of attention, our model is able to draw global dependencies between arbitrary positions in a sentence and yield more interpretable models while allowing for significantly more parallelization. Experiments on open CWS datasets show our model to be competitive in performance and speed. The research reminds us of the potential and trends of neural segmentaion model in the future.

Keywords: Natural Language Processing, Chinese Word Segmentation, Machine Learning, Language Model, Neural Network, CNN, RNN, Attention Mechanism.

目 录

第 1 章 绪论.

1.1 论文选题的背景与意义

语言是人类区别其他动物的本质特性。在所有生物中，只有人类才具有语言能力。而且人类的许多其它智力和能力都跟语言的出现有着密不可分的关联。甚至可以说如果语言没有出现，人类的思维也就无从谈起，知识也无法传承。而近年来，随着大数据时代的来临，互联网上的数据量急速膨胀，而其中一大部分都是以文本形式表示和储存的数据。怎样让机器代替人来快速地处理海量文本甚至语言数据，并从中准确地自动分析和提取出有用的信息，这是自然语言处理领域的所要研究解决的重要问题。而另一方面，处理自然语言的能力是人工智能的一个核心组成部分。概括地说，自然语言处理就是用计算机来处理、理解以及运用人类语言，它属于人工智能的一个分支，是计算机科学与语言学的交叉学科，又常被称为计算语言学。

而几乎在所有中文自然语言处理任务中，一个基本但十分重要的前置工作便是进行自动分词。不仅仅对于汉语文字，对世界上大部分的语言来说，词都被认为是粒度最小但有意义且可以自由使用的语言单位。在口语中，使用同种语言的对话者之间能通过语音的音节、语气甚至声调的抑扬顿挫等分辨其中的词语或短语。而在书面文字中，中文等东亚语言通常由连续的字序列组成，并没有像印欧语言那样的词与词之间天然的边界或分隔符。因此，大部分能够处理汉语语言的人工智能系统都是建立在分辨或识别语言文字中的“词”并将其提取出来的基础上，被称为“中文自动分词”任务。一般意义上来说，自动分词就是由机器在应该成词的文本片段之间自动加上分隔符，从而将连续的中文文本切分为词语序列的过程。

本文在神经网络逐步发展和深度学习技术日益成熟的背景下，创新地提出使用语言模型和注意力机制结合进行自动分词的方案，很好地解决了句子中中长距离语义依赖的问题，并取得了分词性能和速度上的提升。不依赖于现有流行的神经网络分词方法中常用的循环神经网络和卷积神经网络，大幅增加了并行度且减少了计算量，同时保证了分词效果。借助神经网络强大的学习能力，使分词结果面对具体自然语言处理任务和不同的领域语料呈现自适应的特征，并最终构建一种不同于以往分词模式的高效且实用准确的分词模型，对神经网络在分词领域的进一步实际应用具有相当重要的意义。

1.2 国内外研究现状

从中文分词研究早期到现在，出现四大难题是：（1）对于“词”清晰的界定；（2）“分词”与“理解”的先后问题；（3）分词歧义消解问题；（4）未登录词(简称 OOV)的识别问题。下面我们对这几种问题进行详细的介绍和分析。

之所以要明确词的定义，因为在利用机器进行中文自动分词的其中一个至关重要的假设便是：我们能够清楚、统一地界定真实中文文本中每个汉语词语的边界，也就是说我们对汉语词语本身首先要有一个明确的、有计算意义的定义。然而，对于汉语文字来说构建一个统一、清晰且严格的“词”概念的难度相当巨大。根据调查研究，即便我们对母语为汉语的大量人人群做实验，中文词语的平均认同率也只有 0.76 左右[1]。而随意翻开一本汉语语法教科书中，都可以找到有关“词”的一条非常抽象的类似如下的定义：语言中有意义的能单说或用来造句的最小单位。然而从可计算的角度来说，我们认为这种模棱两可的定义是不可计算的，或者说，是不可操作的。早在 1993 年，中文分词研究的初期，为了能明确地给汉语词语下一个明确、统一且完整的定义，经过语言学界和信息学界的等一众研究人员的共同努力之下，《信息处理用现代汉语分词规范》[2]作为国家标准公布，这一标准按词类分别给出了各类分词单位的定义，同时为了和语言学中更严格的“词”概念有所区别，在“规范”中汉语文本中的词语被命名为“分词单位”。然而，令人无可奈何的是，在这一规范中的一些地方，不可避免地把“结合紧密、使用稳定”视为分词单位的界定准则。众所周知，像“稳定”和“紧密”这样的判断其实仁者见仁，智者见智，并不客观。因此，在当时这一规范给分词系统的实现还有评测工作上都给人造成了相当大的困扰和疑惑。然而在当时有人认为“词表+分词规范”的标准应该可以更好地界定句子或文本中的汉语词语，亦即经过在大规模语料中进行统计分析和人工筛选，构建一个统一词表，定量化地实现“使用稳定”和“结合紧密”[3]。而在这段期间国内举办的数届 863 和 973 等分词评测比赛[4][5]也是依照该统一分词规范的思路来组织进行的。这些评测的主要组织思路如下，主办方不提前公布与测评相关的词表文件或分词语料，参赛者直接提交其研发的分词系统后由主办方对系统的分词性能进行测评。而且在测评时，对于参评系统分词输出结果还具有一定程度上的的“柔性”判别标准。亦即待测系统的分词输出与标准答案之间就算有不一样之处，但如果分词输出没有违反“结合紧密，使用稳定”的这一规范，那么这一与标准答案的差异之处就不算错误。然而这一评测方法的不足之处在于评价是否“紧密”和“稳定”需要在一定程

度上引入评测人员的主观判断，而没有一个客观统一，无人工干预的标准。比如分词评测中的召回率（Recall）指标，其分母本来应该是标准答案中的总词数，但在引入柔性化的准则之后，召回率的分母究竟是标准答案中的总词数，还是参评系统输出的分词结果中符合“柔性”答案的总词数？前者相当于直接弃用了引入的“柔性”准则，因为其忽略了标准答案的不稳定性所带来的偏差；后者则又在一定程度上损失了可比性，使得评测本身显得不那么公平客观。出现这一两难局面的根本原因还是在于前面所提到的，缺少对词语本身的清晰界定，即可计算定义。

而 SIGHAN Bakeoff[6]比赛的出现一定程度上解决了中文词语可计算的定义的问题。SIGHAN 的英文全称为“Special Interest Group for Chinese Language Processing of the Association for Computational Linguistics”，又可以理解为“SIG 汉”或是“SIG 漢”。而 Bakeoff 则是 SIGHAN 所主办的国际中文语言处理竞赛，其主要目的与之前的测评比赛相反，不是为了制定统一的分词规范，而是推动中文自动分词技术的整体进步。也就是说，Bakeoff 竞赛的意义和创新在于，它认识到在可以预见的时期内，学术各界几乎不可能对分词标准达成一种共识。那么我们退而求其次，提供了一个公平客观的平台，让多个不同标准的分词语料共同检验分词系统的性能。此外，每种语料都经过各自单位的人工审定，保证了至少在单个语料库内部的分词标准一致性，从而使得参赛系统可以根据这一标准进行学习。许多年来，难以根据一个公认准则判断的某些技术和系统优劣的问题，现在可以通过相同的测试环境下的检验结果进行判别评价。简而言之，Bakeoff 比赛通过分词规范、词表、以及分词语料库三者有机结合的评判准则，使中文词语在真实文本中某种程度上获得了可计算的定义，这也正是实现计算机自动中文分词的基础，和进行具有可比性的分词评测的前提条件。

另一个摆在中文分词面前的难题是分词和理解孰先孰后的问题。这一问题被提出是因为自动分词处于中文自然语言处理领域体系中的底层，其被认为是大部分中文信息处理系统的前置操作或者说是第一道工序，是对句子实施词法、句法以及语义分析的前提。也就是说，中文分词所能依据的只有文本的表层信息，一般是字级别的特征信息。所以，尽管人在识别句子中的词语时是在理解的基础上进行的，人们对词语进行正确的切分取决于是否能对词语所在文本的语义有正确理解和把握。然而从实用角度考虑和可实现的层面出发，计算机自动分词系统不可能完全仿照人类大脑中分词和理解同时进行的模式进行处理，因此大部分研究人员通常会选择“分词在先理解在后”的处理顺序和逻辑。

然而另一些研究人员秉持着对自然语言的理解是一切文本自动分析任务，包括中文分词,的基础, 所以提出了一条截然相反的技术路线——“理解在先分词在后”。把这种处理思路认真加以实现的研究人员之一便是 Wu。他在文献[7]重点分析研究了分词中的消解歧义问题。其主要思想是把中文分词与其构建的句法分析器 NLPwin 进行有机结合。NLPwin 是一个基于语义和句法规则以及策略，且功能十分强大的句子分析系统。该分析系统的词典的开发工作是基于北京大学的一个语法信息词典完成的，当时该词典中单单词条数目已扩充至 8 万条以上，此外还嵌入了大量规则和策略等先验信息以实现歧义消解的功能。他提出把分词的决定放在句法分析的过程中去解决，使得分词和句法分析交替进行，而不是过早地在句法分析前就做出分词的决定。Wu 的一个观点是，传统的最大匹配分词算法(Maximum Matching，简称 MM)缺少利用全局信息的机制，而基于统计的方法则缺乏句子级别的结构信息。句法分析器可以同时提供这两个层面的信息，因此可以在“理解”整个句子的语义和句法的基础上达到最佳的消歧义效果。但当时还没有出现一个公平公开且客观可比的分词评测语料，Wu 在一个自选的小测试集上完成了其研究的分词器效果评测的实验。尽管在这一小测试集上，Wu 的分词方法取得了很高的准确率。但在 Bakeoff 数据集上的的测评结果显示，分词效果并不十分理想，因此 Wu 本人也不得不承认引入句法分析器实际上对分词性能的提升在大部分情况中是十分有限。此外，在基于另一个大规模语料的封闭设置测试实验中，甚至出现了加入句法分析的分词系统的性能低于不使用句法分析的分词系统的尴尬结果。最终出于实用目的，Wu 本人也同意在分词中将放弃与句法分析结合。因此，目前大多数的自然语言处理研究学者和中文分词的主流方法都默认采用“先分词后理解”的方案。

但分词和理解是否只能以两种孤立的过程进行？是否能将分词和理解两者的先后关系打破？换言之，是否能将分词和理解置于同一个模型内，同时进行，并在模型不断迭代的过程中用一方的结果优化另一方的结果，即实现“分词和理解的互动”？由于当时计算能力和语料库规模的限制，采用“先理解后分词”的方案并不太现实。而近年来深度学习技术越来越成熟，计算机的处理能力也得到大大提升，同时面向各种领域、不同任务的语料库的规模也以几何级数的形式增长。怎样运用深度学习的技术，利用多层的神经网络，将“分词”和“理解”或是自然语言研究领域的其他具体任务有机地结合起来，同时提升分词的效果和具体任务的结果，是目前的一个值得思考和探讨的问题。

在学界对上述两个问题的讨论逐渐达成共识之后，剩下两个亟待解决，同时也是各种中文分词方法都不能完全避免的问题便是分词消歧和未登录词识别。歧义是存在于各

种语言中的各个层级的语言单位的一种现象。不仅仅是对机器，对人类来说，同样一个单词、短语或者句子在不同的语境下显然就可能有多种理解，具体表示为多种可能的语义。而分词消歧就是要在自动分词的过程中处理词语级别的歧义问题。例如，“乒乓球拍卖完了”，不同的词语切分方案会使句子有截然不同的意思：“乒乓球|拍卖|完了”或者是“乒乓球拍|卖完了”。两种分词方式对后续语义提取的处理显然有相当重要的影响，而这种歧义的消解往往又需要参考上下文的语境。因为在一般情况下，歧义都是可以根据相应的语境和场景的规定等先验知识而得到解决的。而这也就是我们平时在使用自然语言的大部分时候并不会感到歧义的存在，和能用自然语言进行正确交流和阅读的原因。从这一角度，我们又看出分词和理解密不可分的关系。分词和理解的先后关系和歧义消解两者之间相互依存、密不可分，分词消解歧义的难度由此可见一斑。

而相比于词级别的消歧，未登录词的处理显得更加棘手。由于中文语言词集是一个不断发展变化的开放集合，现阶段不存在任何一个词典或统计方法可以尽数列出所有的汉语词。因此，当分词系统处理不同领域的语料时，常常因为缺乏领域相关的特殊词汇或专有名词的参考，造成分词系统产生错误的切分。例如，根据在 Bakeoff 公布的测评数据汇总后进行的统计估算结果显示，仅仅因为未能成功识别未登录词造成的分词精度下降程度至少比未消解分词歧义造成的损失至少要大五倍以上[28]。由此可见，如果一个分词方法识别未登录词的能力相比现有水平能有大幅度提高，那么势必将带动整体分词效果的显著改善。

为解决未登录词的问题，传统分词系统一般的做法是通过加强专有词汇的搜集来补充完善特殊领域的词典，并在该词典基础上实现最大匹配法的机械分词。因此自动抽取新词汇或关键词的算法成为这些分词系统的先期准备步骤甚至决定最终分词效果的核心技术。领域内的专有关键词汇多出现在该领域相关的文本中而鲜有出现在其它领域中，因此大多数关键词抽取算法都基于此特性进行设计。一般来说，高频关键词比较容易抽取，而少数低频的新词难以自动地进行事先搜集，必须结合在线辨识进行提取。常用的共现信息包括词素、构词律、词汇及词汇间的共现讯息，为在线新词等未登录词提供辨识抽取的依据。

而目前主流的中文分词工具则利用隐马尔科夫模型（Hidden Markor Model, HMM）甚至条件随机场（Conditional Random Field, CRF）等序列标注概率模型来识别未登录词。这种方法将分词转化为序列标注问题处理，通过给每一个字打上词位标记来确定词的边界，在实际应用中能很好地处理未登录词识别。这类处理方法虽不需要构建分词词典，

但是需要大量分好词的训练语料用以学习，因此在一般使用中还是与传统的词典和规则匹配法进行句子切分。总之，目前来看，仍然没有一种分词方案能在未登录词识别上给出两全其美的解决方案。

中文分词发展早期遇到的两个问题：“词”是否有清晰的界定以及分词和理解到底应该以怎样的顺序进行的问题。当时受限于硬件运算性能和语料库规模，分词任务需要清楚地定义词与词间的边界，或者通过分词规范、分词语料库和词表三者相结合使分词标准化，一定程度上使中文词语得到了可计算的定义，便于对比不同分词方法之间的准确度。但这种分词的标准化在实际应用中是否确实有不可或缺的作用？固定的分词结果是否能适应不同领域的自然语言任务的实际要求？是否能使分词结果具有一定的灵活性，能根据实际任务的需求，不断地自动调整词语序列的组成方式？这是在目前中文分词准确率在一定程度上已经到达瓶颈的情况下，十分值得思考的一个问题。

同时也有人提出“先理解后分词”的思路，通过将中文分词建立在句法分析或词性标注任务的基础上，借由这些方式，在初步“理解”内容的同时进行分词，以优化分词的结果。但这同时也增加了分词的复杂度，损失了速度，而分词精度也没有可观的提升。出于实用性考虑，早期学者大多默认采用“先分词后理解”的方式处理自然语言。而近几年来，随着人们对自然语言处理的不断深入研究，各种不同类型的语料库也不断充实，一些先进的分词系统开始尝试加入句法分析和词性标注所提取的信息以提高分词精度，取得了相当不错的成果。但在深度学习领域这种分词方法却没有受到广泛关注。一是因为基于深度学习的分词技术起步较晚，尚没有任何一种神经网络本的分词方法取得工业级的应用；二是由于句法分析需要人工设计特征，与深度学习自动学习特征的思想相悖。而词性标注的语料库大多独立于分词语料库，用两种截然不同的语料库对分词神经网络模型的训练难以实现。但无论如何，若能将利用更高层级的信息，如短语级别、句子级别甚至篇章级别的信息，势必能将分词准确度提推向一个崭新的水平。

1.3 创新点和技术路线

近年来，随着深度学习和高性能计算技术的日益发展成熟，学者们将原本在图像领域获得巨大成功的深度神经网络模型拓展应用于其它领域，自然语言处理便是其中之一。最近几年，神经网络模型在多个自然语言处理任务上媲美甚至超过了最优的传统方法，例如在机器翻译、语音识别等公开数据集上都获得了最先进的成绩。神经网络结构因其

能自动提取特征，免除过去大量花费在特征工程的时间和人力，无需外部的先验知识，只要求少量的人工干预，受到众多研究者的青睐。

与卷积神经网络近期在图像处理领域拥有大规模应用不同，循环神经网络因其具有对时序特征建模的天然结构优势，在自然语言处理领域似乎更受研究人员的青睐，这一点对于中文分词领域来说也不例外。一段时期以来，对于中文分词的最新研究著作主要都集中于利用神经网络，特别是循环神经网络对句子序列进行建模，自动提取对分词有意义的特征信息。包括两种主流的实现思路，基于字序列标注的分词模型和基于词序列最优解搜索的模型。而且研究人员发现循环神经网络的几种变体，例如长短期记忆（LSTM）以及门循环单元（Gated recurrent unit），相对原有的设计能更容易地捕获对序列中长期依赖关系。在分词任务中，这两种循环神经网络的变体应用相当广泛，并发展出双向结构（Bi-directional）和堆栈结构（Stacked）等多种变形，取得不俗的分词效果，而在本文中也将着重分析探讨这几种网络结构在分词上的应用。对于序列标注模型来说，还可在神经网络提取特征的基础上，加入条件随机场等统计标注模型，学习状态转移概率矩阵，从而获得更高精度的分词器。在包括 Bakeoff 在内的多个公开数据集上，这些基于神经网络的分词模型都取得了足以和传统最优分词系统相媲美的测评成绩。但令人遗憾的是，在中文分词领域，最新的神经网络结构并没有像和其他自然语言处理领域一样，大幅领先于传统方法。这在一定程度取决于分词本身处于自然语言处理任务的底层，直接提取特征建模的难度较大，也和分词问题的提出已有相当长的历史，各类理论研究相对扎实且完备和实际应用也已逐渐完善等因素有关，因此基于神经网络的新分词方法的研究在短时间内还很难超过已经过充分优化的传统分词系统。然而从这些新提出的分词方法本身来看，大部分模型还是在已有的基于字序列标注或词序列搜索的方案基础上套用循环神经网络结构，尚缺乏对特定分词任务构建专门的深度神经网络结构的探索实践，而从深度学习在其它领域的最新成果来看，只有针对实际任务设计相适应的网络结构，才能最大化地发挥神经网络自动特征提取的能力。

另一种常用于对序列建模的深度学习方法便是注意力机制，在编码解码模型（Encoder-Decoder）中采用注意力模型使得输出序列任意位置可直接获取输入序列任意位置的信息，从而获得较好的对长距离依赖关系建模的能力。而在一般的卷积神经网络或循环神经网络中，通常需要经过数次的网络层传递才能使输出位置关注任意输入位置的权重向量。正因如此，2017 年来自谷歌的团队仅仅使用纯注意力建模的深度神经网络，在两个德语——英语，法语——英语机器翻译权威数据集上获得了最高的测评分数。因

此，注意力模型也收到了更加广泛的讨论和研究，被使用于各个领域的深度学习模型，特别是序列转导模型中。本文的最大的创新点在于首次将自我注意力机制运用于中文分词中，完全不依赖于卷积神经网络和循环神经网络及其变体，并针对分词任务目标的特点设计了多层自我注意力单元对句子序列建模，在不增加了网络整体计算量的前提下大大提高了神经网络分词模型的并行度。模型分词效率较高的同时，不借助任何外部语料或知识图谱，在多个公开数据集上取得了与先进分词方法相当的测评结果。

本文将在后续的章节中探讨在一些中文自动分词领域的最新研究成果中，直接将循环神经网络结构套用于分词任务的不足之处，并引出同样适用于序列建模且计算并行度更高的注意力模型。论文中的一大创新点在于第一次将多层自我注意力模型运用于中文分词任务中，并且不依赖卷积神经网络或循环神经网络，也不借助于外部语料或先验知识，仅靠注意力原理搭建的网络便能很好地对句子序列中任意位置之间依赖或权重关系建模，自动提取出有利于分词的特征向量。从而表明深度神经网络在中文分词领域可能具有进一步发展的方向和潜力。

1.4 本章小结

在介绍了国内外的分词研究现状和论文的主要创新点之后，本文先从将文本中的字符序列转化为计算机可理解的结构化表示的技术开始，介绍几种先进的统计分词技术和将循环神经网络应用于中文分词的最新模型，比较这几者之间在分词速度和精度上的差异。然后引出注意力机制，介绍其原理，并探讨其为何能对时序输入中的长距离语义关联建模。最后阐述笔者提出的注意力分词网络结构，并介绍对比实验中各个超参数设置所带来的影响，以及在各种公开语料库的封闭测试集的分词结果。

第 2 章 文本分布表示技术与中文分词模型

数据表示是机器学习中，同时也是自然语言处理中的一个基础工作，数据表示的好坏直接影响到整个后续任务处理的性能。在把数据，如语言文字交由计算机进行自动处理之前，我们须要将其从符号化表示的的原始文本序列转化为一种结构化的，可直接被机器储存传递计算的形式。如何针对具体任务，设计一种合适的数据表示方法，以提升机器学习系统的性能的工作也被称作特征提取或者说特征工程。然而，对于自然语言来说，并没有像图像或语音数字信号那样直接可用的表示方法，这也为给自然语言的特征提取提出了更大的挑战。

在 1954 年分布假说的出现为构建文本的结构化表示提供一个十分重要的思路，其基本假设是：上下文相似的词，其语义也相似[56]。而在 1957 年，Firth 其著作中进一步明确和阐述了分布假说，他提出的一个假设推论是：词的语义由其上下文决定[57]。换一种说法，分布假说指出我们应该关注两种对象：词和上下文，而当前词的上下文决定了该词的语义。基于分布假说，后来研究自然语言处理的学者们提出了多种词表示模型：如基于矩阵的 LSA 模型[58]、基于聚类的 Brown clustering 模型[59]以及基于神经网络的词表示模型。这些模型的核心其实主要由两个基本要素组成：一是要确定当前词的上下文的范围并选择合理的形式定义或代表上下文；二、选择一种模型描述某个词或当前词与其上下文之间复杂多变的关系同时使其与分布假说相符。而不同文本表示模型之间的最大区别便在于如何对当前词和上下文间的联系进行建模。

由于词是语义的基本单元，因此一般自然语言处理任务中将词的向量表示作为模型的输入。然而中文分词的目标就是将字符序列切分成词语序列。因此除基于词典的匹配分词方法外，绝大部分中文分词模型从更底层的单元，即从字的向量表示出发，从这些特征向量中挖掘词的信息。由于词的向量表示技术都适用于字的向量表示，因此在本章中，我们将研究探讨这些表示方法。

2.1 传统的向量表示技术

2.1.1 基于词频的向量表示

自然语言处理中最直观、最简单，也是到目前为止最常用的文本表示技术是独热表示，这种表示方法把每个字符或词语与一个值为布尔值且维度大的稀疏向量。该向量的维度大小即对应于词表大小，亦即收录在词表中所有字符或词语的总数，只有对应该字

符或词语序号的维度上值为一，而其余维度上的值为零。例如:“你”表示为 [1 0 0 0 0 …]，“好”表示为 [0 1 0 0 0 …]。如果我们用一个一维整数数组表示一个文本片段，对于每个字或词我们可以用其在独热向量中的序号这种非常简洁的计数方法来表示，例如“你”记为 0，“好”记为 1。

基于这种方法，我们可以通过将每个字或词的独热向量简单相加构建上下文表示，这种表示技术亦被称为词袋（bag of words, 简称 bow）模型，该模型忽略掉上下文中的句法和语序等结构要素，将其仅仅看作是若干个字符或词汇的简单集合，即假设上下文中每个单词的出现都是相互独立，不存在依赖关系的。词袋模型使用一组无序的离散稀疏向量来表达字或词的上下文。该向量的定义方式与原来在文本中各个单词出现的顺序完全没有关系，而向量维度上的值等于该维度所对应编号的字符或词语在整个文本中出现的频次。

虽然这种方法十分简洁而且直观，然而对于两个意义相近的词语，如“男”和“女”，使用独热表示的话我们会得到两个完全不同的向量，但事实上很多时候这两个词是语义相近甚至可以相互替换的。其次，随着语料的增加，词语的数量越来越多，向量的维度越来越冗长，很容易遇到维度灾难的问题。在一般的自然语言处理任务中词汇表收录的总词数一般都非常多，甚至可以达到百万的量级，如果每个词都对应地用百万维的向量来表示可能使得机器难以继续处理，所以这一现象也被称为“维度灾难”。后文提到的基于神经网络的分布式表示技术则很好地解决了这一点。

2.1.2 基于矩阵的向量表示

基于矩阵的向量表示是基于词频方法的一般化形式，也可以认为其是基于词频表示的一种拓展和延伸。这类表示方法的主要思想是构建一个词与上下文的矩阵，从矩阵的每一行中就可以提取对应词的向量表示。不同的基于矩阵分布表示的区别在于上下文的选取和计算矩阵中元素值的方法。一般将词所在的文档或者词附近上下文中的各个词或 n 元词组（n-grams）的词袋表示作为上下文，并将原始共现次数经过加权和平滑后的值作为矩阵的元素值，以提升模型的效果。常用的算法有 TF-IDF、PMI 和直接取 log。

最常用来计算矩阵元素值加权算法的便是词频-逆文档频率法（TF-IDF）[29]，这一方法是由 Salton 在 1988 年提出的，它是一种在词袋模型的基础上进行拓展的表示方法。TF-IDF 算法是基于这样一个非常合理且实用的统计学规律之上：对区别不同类别最有参考价值或者说最有辨别力的词语，一般情况下是那些在小部分特定文档中出现频次高，

但在其他文档甚至整个文档集合中出现频次相对较低的词语，而这一理念与前文提到的构建特定领域词典所用的新词提取算法有一些相似之处：领域新词通常是那些在专有语料中出现频率高，而在通用语料上出现频率低的词。总的来说，该模型主要统计两部分信息：

每个词的词频，也就是词语在文档中重复出现次数和与该文档中包含重复词的总词数的比值。除以总次数的意义在于对出现频数进行归一化处理，防止词频权重偏向较长的文件，从而得到词语在该篇文本上的离散概率分布。

每个词的逆向文档频率，即是指文档集合中文档总篇数与包含该词的文档篇数的比值，再取对数。相对于一般的词袋模型，加入逆文档频率信息的主要目的是抑制文档内无意义词或高频词对分析判断文档类别或属性带来的负面影响。因为一般来说出现频率最高的词就是在自然语言处理中常提到的停用词，例如“的”、“是”等，其对应的权重应适当降低。所以如果一个词在文档中越常见，逆文档频率就越接近于 0。

最后所得的 TF-IDF 权重即为这两部分的乘积：

$$\mathrm{TF-IDF}=\mathrm{TF}*\mathrm{IDF}=\frac{d}{D}*\ln\left(\frac{n}{N+1}\right)$$，D为词语出现次数，N 为文章总数

可以在生成的“词-上下文”矩阵的基础上，使用包括 PCA 在内的一些数据降维技术将这一高维稀疏向量压缩成低维稠密向量，从而减少噪声带来的负面影响。

2.1.3 基于聚类的向量表示

基于聚类的向量分布表示又称作分布聚类（Distributional Clustering），这类方法通过无监督聚类手段构建词与上下文之间的关系。其中最经典的方法是布朗聚类[30]。布朗聚类属于层级聚类方法，其可以构建一个多层聚类体系，对每个词的预测结果为该词的对每个层次中每一类的条件似然。因此可以根据两个词所属的公共类别与类别总数的比值判断这两个词的语义相似度。具体而言，布朗聚类的学习目标为最大化以下似然估计，其中c_i为词w_i对应的类别：

$$P(w_i|w_{i-1})=P(w_i|c_i)P(c_i|c_{i-1})$$

布朗聚类只是简单地考虑了相邻词之间的关系，也就是说，每个词只使用它的上一个词，作为上下文信息，对应于二元语言模型。

2.2 基于神经网络的向量表示

2.2.1 神经概率语言模型

基于神经网络的向量表示技术，由于可以将语义信息融入到向量中，所以又被称作基于语义的向量表达。这一技术的提出与语言模型（Language Model）[31]密不可分。语言模型可以对一段文本是否为自然语言的概率或似然进行估计，即用概率来表示其文本片段真实存在的可能性。对于一个$w_1 w_2 \dots w_n$组成的字符串，通常采用下式计算其概率值 P：

$$\mathrm{P}(w_1 w_2 \dots w_n) = \mathrm{P}(w_1) * P(w_2|w_1) * \dots * P(w_n|w_{n-1} \dots w_1)$$

因此常用的自然语言模型即为求$P(w_t|w_{t-1} \dots w_1)$的概率，而 n-gram 模型则用来近似$P(w_t|w_{t-1} \dots w_{t-n+1})$代替该概率。

自本世纪初开始，就有大量基于二元语言模型，利用神经网络建模的研究工作发表。而其中最经典的模型要数神经网络语言模型（简称 NNLM），它是由 Bengio 等人在 2001 年发表的论文[32]中被首次提出的。该模型使用一个三层前馈神经网络来构建语言模型，这三层网络分别被称为输入层、隐藏层和输出层。

其后许多人围绕着优化算法、节约计算量以及重新设计模型结构进行了一系列后续工作，例如 C&W 的 SENNA[33][34]（Word Embedding），M&H 的 HLBL（Hierarchical Log-Bilinear）[35][36]用层级思想替换了矩阵乘法，极大提升了速度，Huang 的语义强化[38]（解决了多义词问题）等。而其中最具有代表性的成果则要数谷歌 Tomas Mikolov 团队的循环神经网络语言模型（RNNLM）[37]。

2.2.2 Word2Vec 词向量模型

谷歌团队于 2013 年发表的 Word2Vec[39][40]无疑是至今为止最流行的获取词向量表示的开源工具包。一经面世就因其模型结构简单、算法高效的特点，引起了众多学者和研究人员的关注，其后相当一部分自然语言处理研究工作都借助 Word2Vec 工具预训练词向量，用作网络输入嵌入层的初始值并进行 fine-tune，相比原有模型得到不少性能上的提升。

Word2Vec 的主要原理是利用浅层的前馈神经网络对词与上下文或词与词之间的关系进行建模，预测每个词对于上下文中相邻的词的共现概率，实现对原始高维稀疏离散向量进行降维。上下文的范围通常是以当前词为中心，左右长度相等的文本片段窗口。词向量神经网络训练完成后，其中的隐藏层矩阵将词表中的每个词映射为低维、稠密、

实数向量，这样得到的词表示一定程度上保留了语义信息，因为对于语义相近的词来说，其词向量表示一般在向量空间上也相近。我们可以使用例如欧式距离、余弦相似度等距离度量来计算这种相近的关系。富含语义的低维稠密词向量不仅能用作深度学习中网络嵌入层初始值，还可以用来自动联系关键词，对一些基础的自然语言处理任务，例如构建知识图谱，作为文本分类的种子关键词等都有十分重要的作用。

2.2.3 Skip-gram 模型和 CBOW 模型

其实在 Word2Vec 中，同时包含了两种神经网络结构的实现：CBOW 模型和 Skip-Gram 模型，网络结构如图 2-1 所示。这两种模型在根据前人经验成果基础上，简化现有模型，保留核心部分。主要体现为去除了原有三层网络结构中的隐藏层，相当于从神经网络结构直接转化为对数线性结构，与逻辑回归一致；并且精简输入上下文表示的构建方法，使得上下文表示向量维度与词向量维度一致。两个模型结构基本相同，建模目标也都可概括为从上下文预测目标词[64]，不同之处在于前者使用上下文窗口中除目标词，亦即中间词外的词向量加和表示上下文，而后者则更简单，每次预测直接选择上下文窗口的一个词作为上下文表示。此外两种模型还采用几种优化方法来加速计算条件概率值和设计更好的目标函数等，如层级 softmax[60]和负采样[61]技术。层级 softmax 可以概括为通过构建一颗叶子节点为词表中的每个词，节点权重为词在语料中的出现频数，且所有叶子节点带权路径长度最小的最优二叉树，即霍夫曼树，使内部节点代表输出层的神经元，并将原始多分类输出的必需的复杂 softmax 计算转化为若干次递归地计算二分类输出的 sigmoid 值的优化算法。从原本每次预测需要与词表大小线性相关的运算次数，优化为运算次数仅为词表大小的对数。大大减少了预测输出目标所需的计算量，降低了时间复杂度。而负采样技术则在提升效率的同时，构造出了一个新的优化目标，期望模型可以最大化正样本的似然，同时最小化负样本的似然。

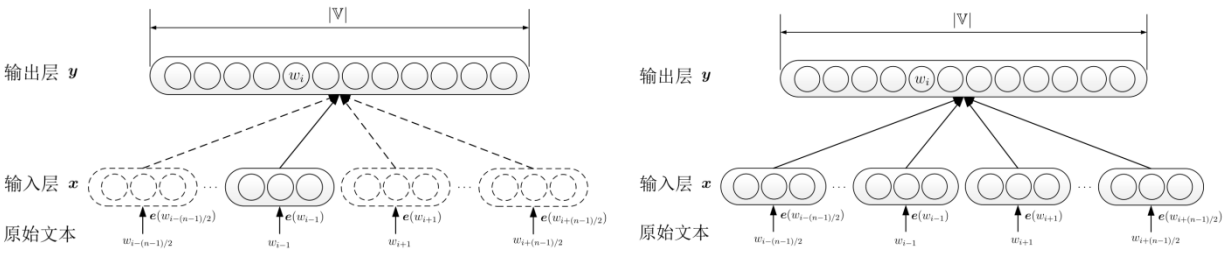

图 2-1 Skip-gram 模型（左图）和 CBOW 模型（右图）结构示意图

2.2.4 训练 Word2Vec 词向量模型超参数设置与调优

由于采用了较多的优化方法，同时神经网络本身也有许多超参数需要调节，并且模型属于无监督学习，没有统一的标准评价模型训练的好坏与否。因此在刚开始实际应用 Word2Vec 时，往往难以找到合适的实验设置和超参数方法，使得模型选择和调优无从下手。笔者将以往实验中调节各类超参数的相关经验和实际效果总结如下：

词向量维度大小：维度越大，一般来说词向量效果越好，因为其中蕴含的信息量也越大。但是模型训练和预测花费的时间也越长，需要做权衡。通常，更大语料规模需要用更高维的词向量来学习。但是使用越高维的词向量，相同词间的余弦相似度会越低；

词窗口大小：窗口越大，上下文代表的范围越大，可能使中间词的语义更丰富，但也可能引入更多的噪声，使得高频词的权重更大。由于模型输入构造方式不同，skipgram 模型训练时间基本与词窗口大小成正比，而 CBOW 模型几乎不受影响；

最小词频：意即如果某个词在整个语料中出现频次低于该数值，则在语料中去掉该词，不训练该词的词向量。低频词的词向量学习难以取得较好的效果，可将其视作噪声去除，一般取一个不高于于 10 的值。由于低频词在语料中的总量少，去除后使得模型总参数变少，softmax 层计算量也减少，而且可以大大减少模型保存文件大小或内存大小；

高频词降采样阈值：意即若某词在语料中的频率高于该阈值，则开始对该词进行降采样，即在语料以一定概率忽略或者说跳过该词，且跳过概率随词频增大而变大。该参数同样会影响高频词的权重，改变样本词分布，且实验结果显示合理调节可改善词向量训练的效果，因为在某种程度上，高频词和停用词的表现很类似。显而易见，调低该阈值会明显加快训练速度；

层级 softmax：主要通过按词频大小构建 Huffman 树，自顶向下预测每个内部节点的概率，减少普通 softmax 层计算量，可以提升模型训练速度。但不能与负采样同时使用，若不需要负采样，则一般可以使用层级 softmax 以降低时间复杂度；

负采样词数：即为对每个上下文-中间词（或称目标词）对，按特定词分布采样负样本的数目，随机采样的词组合一般不具有自然语言的意义，因此可作为负样本。增大负样本词数一般不会显著增加训练时间，同时可以有效提升模型性能；

语料迭代训练次数：一般语料规模越大，迭代轮次可以适当减少。一般可以通过观察训练时目标函数值是否收敛确定迭代终止的轮次，也可根据具体任务构造的验证集上模型的表现来确定，或者通过设置一个较大的值来保证模型收敛。

总的来说，基于神经网络的向量表示技术是建立在语言模型的假设之上的，其主要目的将某种语言中的每一个字符或词语映射成一个固定长度的向量，从而形成一个包含语义的向量空间，我们可以通过某种距离度量来衡量向量所对应的语言符号之间在词法、语义上的相似性。在这一具有语义的词词向量基础上，我们就可以构建更深层的神经网络结构从而对更复杂的上下文依赖关系甚至句法、语法等结构信息深入建模，进一步提取出更丰富的特征和信息。

2.3 基于统计学习的中文自动分词模型

在中文分词研究的初期，一般使用一种被称为机械分词的方法实现中文文本的自动切分。可以理解为，它是按照一定的策略和规则，将待分析的汉字字符串与一个“充分大的”机器词典中的词条进行逐个比对，若在词典中找到与文本片段完全相符的词条，则词语匹配成功，识别出一个词并切分，否则在比对了词典中所有词条都不匹配的话则比对失败，不予切分。在具体实现上，最大匹配法需要将词典储存在一个经过排序的特殊树结构中，以提高查找词典的效率，同时实现高效的多模匹配算法，常用的有 AC 自动机等。

机械切分方法除了最大匹配法，还有最少词数法，最短路径法[44]等。在此基础上又发展出基于词典的概率分词模型，如基于有向无环图的最大概率组合，还有基于语言模型的最大概率组合等等。这些基于词典的分词方法引入了动态规划思想，实现简单算法高效，能较好地解决交集型歧义问题，但对于未登录词的识别率还是较低。一般性能较好的传统分词系统，都是把此类基本的机械分词作为一种粗分手段提供数量较少但召回率较高的不完全的中间结果，还需进一步通过专门的歧义消解，以及识别未登录词的算法或规则来进一步优化文本切分准确率。

与基于词典的机械分词思路截然相反，也有人提出基于统计的无词典分词法，或者说是一种基于领域内文本的新词发现和抽取算法。其基本原理是根据语料库的统计信息判断是否成词，并通过制定阈值输出可能错误切分概率最低的分词结果。具体来说，统计的信息包括词频和字间邻接关系以及字共现关系，例如凝固度（互信息熵）和自由度（边界熵）[62]等指标，并经过人工过滤或设置固定的指标来确定一个文本片段是否是一个词，相当于通过文本挖掘和信息提取等手段构建出一个领域内专有词典，再利用基本分词方法进行句子切分。这种方法虽然领域自适应性较强，能很好地识别领域内中高

频的未登录词，出现歧义的情况较少，运行速度也较快。但是分全率和准确率较低，受语料规模大小限制影响很大，较小的语料很难保证统计结果符合一般自然语言的特征。

在某种程度上，基于词典的基本分词法和基于统计的无词典分词法是优势互补的。因此也有人提出将字典与统计信息结合起来的分词方法[45]。主要实现思路是首先利用基于字典的基本分词方法进行粗分词处理，接着利用语料的统计信息处理歧义问题并识别未登录词，并扩充已有词典。这类方法优化了出现交集型歧义的情况下切分的准确率，并且在一定条件下能很好地识别特定语境中高频未登录词。因此在进行专门领域语料的分词实践中，一般都会沿用词典匹配和新词发现相结合的模式，较高效且准确地进行语料切分。

以上提到的都是基于“先分词后理解”的思路实现的算法，另外有学者提出“分词和理解互动”的新模式，被称作基于语法知识和规则的分词法。其基本思想就是利用事先定义好的规则和策略等先验知识进行语义和句法分析的同时进行自动分词。但正如前文提到的，因为语法知识、句法规则的建立是一项十分庞杂的工作，所以这类分词法所能达到的精确度还不能达到令人满意的程度。

随着分词技术的不断发展，基于字标注的统计学习分词模型成为了主流。这是目前在工业界常见的分词方法，其优势在于能很好地解决未登录词识别的的问题，诸如 NLPIR (ICTCLAS), Jieba, LTP, Ansj, Thulac 等广泛应用的主流分词系统都采用了这种方法，相关的统计学习分词模型的研究也相当广泛和充分。其基本思路是将自动分词当做一个文本序列标注（sequence labeling）任务来处理。通过标注每个词的词位信息，由字构词，从而将整个字符序列切分出来。最常用的是 4 词位标签系统：BMES，四个字母分别代表字处于词首，词中，词尾以及单字词。这种方法主要从字的上下文和标注历史中提取特征，而这两种特征既可以由专家人为设计，也可以用神经网络等方法自动学习。由于要处理的对象是字符序列，因此常采用的可处理序列对象的标注模型，主要有隐马尔可夫模型，条件随机场与循环神经网络等，神经网络分词模型于下一小节中介绍。

隐马尔可夫模型（简称 HMM）最早由 Rabinen[8]在 1989 年提出是马尔可夫链的一种，它可以被看做是朴素贝叶斯分类模型的序列化形式。由于对状态序列和观测序列的共现概率建模，它属于生成式模型，而且是一个双重随机过程。这一模型引申出三个经典的基本问题：评估问题，解码问题和学习问题，可分别由前向算法，维特比算法和 Baum-Welch 算法[63]（向前向后算法）解决。

条件随机场（简称 CRF）是给定输入随机变量序列情况下，输出的随机变量的条件概率分布可构成马尔科夫随机场的概率模型，可以看作是最大熵模型[10]的序列化形式。与隐马尔可夫模型不同，它属于判别式模型，同样适合对序列信息建模，而其特点是假设模型输出的随机变量构成马尔可夫随机场。在分词领域常用的是其线性形式，即线性链（linear chain）条件随机场[9]，其形式被简化为一种对数线性模型。且假设某一时序的输出只与前后的两个输出有依赖关系，因此任一时序的输出随机变量与前一时序的输出以及输入序列三者构成最大团。其学习方法通常是极大似然估计或正则化的极大似然估计，且使用随机梯度下降或拟牛顿法等算法进行优化，同样使用维特比算法预测给定输入的条件概率最大的输出序列。而一般的条件随机场可对输出随机变量具有复杂依赖关系的序列建模。它与最大熵马尔可夫模型的区别在于，后者团（clique）的势能函数归一化是局部的，每个特征的类别分布是不均匀的，会导致标注偏置问题；而前者在此基础上做了改进，其概率归一化更加合理，但缺点是可能导致计算量增加。

总的来说，基于字的标注方法能较好地对未登录词进行识别，同时也可处理交集型歧义的问题。而主流分词工具采用基于词典的匹配和基于字标注相结合的方式在实际应用中取得了令人满意的分词效果，分词的同时进行词性标注和句法分析则能进一步提升其性能，因而在工业界得到了广泛应用。

2.4 基于循环神经网络的字标注分词模型

目前，基于线性链条件随机场的 4 标签体系分词方法是传统统计学习标注模型中表现最好的。但状态特征函数和转移特征函数等仍需通过人为地定义特征模板的方式显式地给定，这就需要花费大量的时间精力在特征工程上。而近几年，越来越多的学者利用神经网络能自动提取特征的优势，使用循环神经网络及其各种变体作为输入文本序列的编码层，并将输出的神经元激活值输入最后的条件随机场解码层，取得了先进水平的分词性能，从而成功将神经网络结构应用于分词领域中。例如，2015 年，Chen 等人[13]将自然语言处理领域备受推崇的 LSTM 单元，即循环神经网络的一种变体，应用于分词任务中，并加入了 Dropout 技巧、最大间隔准则构造的目标函数以及对标签依赖关系建模的转移概率矩阵等，并试验了多层堆叠 LSTM 网络等变体。最终在多个公开数据集上取得了领先的测评成绩。其后在 2016 年，Yao 等人[65]进一步提出了双向 LSTM 分词模型，即使用两层对输入序列处理顺序相反的 LSTM 网络，并将其输出拼接后作为下一层

的输入。堆叠多层双向 LSTM 单元后，该模型在 Bakeoff 数据集上的测评结果相对单向形式达到了更高的水准，刷新了同类神经网络分词模型的最好成绩。

而 He 等人[18]在 2018 年提出利用语言学知识，从比“字”更小的语言单位中获取更丰富的语义信息。即通过以无监督的方式联合训练部首与字的向量表示，来增强预训练字向量的表示效果。并以基于字标注的双向 LSTM 分词网络为例，证明了部首向量的加入确实可以提升分词准确率，从而推断这种字向量预训练方式的有效性适用于所有象形文字的序列标注任务。

其基本处理逻辑是将输入到输出节点的状态特征函数通过循环神经网络或层叠的卷积神经网络隐式地定义，而相邻两个标签的转移特征则由一个状态转移概率矩阵直接给出，且该矩阵的权重随训练的进行以一定频率更新，而神经元权重则由目标函数根据梯度下降法进行更新。这样加入转移概率矩阵强制定义输出标签的关联，相对于直接由逐标签 softmax 得到输出序列，好处在于加入了一定的标签转移约束条件，使得编码模型无需直接关注输出层面的上下文关联。整个模型也获得一个较好初始值，用于编码的神经网络权重更容易收敛，另一方面也减轻了人工定义特征模板的负担。实验显示，神经网络结合条件随机场的分词模型在公开数据集上获得了优异的性能表现。

由于此类模型采用了序列标注方法来进行分词，常常被拿来和另外两个同样处于自然语言处理底层且属于序列标注的任务，词性标注（简称 POS）和命名实体识别（简称 NER） 的模型相比较，甚至使用统一的模型对这三种任务同时学习。2013 年，复旦大学的 Zhao 等人[41]便提出一种对中文分词和词性标注同时进行多任务学习的模型，在使用两种基于不同标签体系的人工标注语料库的情况下，他们建立了一种宽松且具有不确定性的标签映射方案。区别于以往模型在多任务学习时通常都以顺序式的结构对异构的语料库进行学习，Zhao 等人提出的统一标注模型可以同时在两种语料库上训练，获取不同标签体系间共享的信息，且在两个数据集都获得了较大性能提升。而 2018 年，来自百度的 Jiao 等人[42]更是通过神经网络结构对中文分词、词性标注、命名实体识别三种序列标注任务结合起来同时训练。以往处理这三种任务时需要结合多种词法分析工具且须人为定义大量规则和策略尽量避免不同工具之间出现词边界冲突和标签冲突，该模型的提出不仅大大减少了解决此类冲突花费的时间精力和计算量，还避免了先后使用不同词法分析工具所造成的错误传递问题，也将不同标签中的共享信息有效地利用起来，且只需要少量的人工标注语料库便显著降低了错误率。

但是对于字标注的统计学习分词方法来说，一个显而易见的缺点便是需要大量的已分词语料训练模型。当分词语料拓展到新领域时，同一模型的表现与已知领域比较可能相去甚远。而标注语料所需的人力又远远大于构建领域新词词典。因此如何将基于词典分词的优势结合在统计学习标注方法中是一个相当重要的问题，在实际应用中，一般在解码的最后阶段加入人工干预，又被称为硬解码。常用做法是根据词频将词典内每个词的对应的字标签输出概率直接进行调整，例如对高频词中每个字输出相应位置标签的概率乘上一个大于 1 的值，而对于错误切分的片段则对相应字标签乘上一个小于 1 的值。

2.5 基于最优解搜索的词级别分词模型

2.5.1 基于最优词序列搜索分词方法的基本原理

上述基于神经网络的字标注模型虽能自动地从字序列中提取特征，但是其难以直接利用词语本身的信息，因此分词效果可能难有进一步的提高，基于最优词序列搜索的分词方法则可以很好地解决以上问题。一般思路是对所有可能出现的句子切分情况进行评分，将得分最高的分词结果作为输出。其优势在于可利用多个级别的信息对分词结果进行评分：词级别，即该词汇单独出现的似然估计；句子级别，将每个词联结起来之后整个句子自然合理的分数，即对语义和句法进行打分。

具体实现方法可以是仿照人的阅读习惯，从左至右地对句子逐步进行切分：将已分词的部分的分数与候选词的词汇分数以及两者结合的分数相加，作为新的已分词部分的打分。这样做便将分词问题转化为树搜索问题，虽然问题的求解空间随句子长度呈几何关系增长，但可以使用剪枝算法优化求解速度，求得局部最优解。而如何将不同数量的字向量表示转化为固定长度的词向量表示是该方法的另一大难点，使用的网络结构太复杂会导致运行时间过长，甚至过拟合；结构太简单则不能充分利用字级别的信息对词进行向量表示。

2.5.2 神经网络词级别分词方法最新研究

近两年关于利用神经网络结构实现基于最优词序列搜索的研究方兴未艾，Cai 等人 2016 年使用 LSTM 网络结构和门单元对该问题建模[12]，如图 2-2 所示，并于 2017 年在不损失准确率的情况下简化了模型结构[23]，显著改进了分词效率。同期，Zhang 等人[19]在 Zhang 和 Clark 等人的研究工作[20][21][22]基础上，提出了基于转移的模型（Transition-based）的神经网络版本，免去了人工构造特征的工作。以上两者的研究都

在利用已分词序列中每个词的语义信息之外，尝试从字的级别乃至分词操作历史或上下文级别中提取对分词有益的特征。此外，他们都采用了打分的方式评价已分词部分的好坏，用剪枝算法来局部搜索最优的分词序列，并选取最大间隔准则来构造训练模型的损失函数，甚至都使用 LSTM 神经网络来从输入中递归地提取上下文全局信息。

两者的不同主要集中在如何选择或构造输入特征上，前者增加了一个通过门控制将字向量组合转换成词向量的网络层，使得字级别的特征传递到了词中，并与另外一个原生的词向量结合共同作为网络的输入；而后者的模型输入为字、词、分词操作三个向量序列，并将其分别输入三个独立的 LSTM 网络进行建模，相比之下模型结构更复杂，参数更多。除此之外，两者在给分词序列打分的依据上也有所不同，前者将 LSTM 隐藏层输出和输入词向量作为判据，分别用来权衡分词后序列上下文是否通顺以及分出的每个词是否合理；而后者将三个 LSTM 层输出前后拼接，经过一个双曲正切激活层后计算其打分，还加入 Zhang 和 Clark 等人在传统方法下构造的离散特征作为额外的分词操作打分判据。另外，前者只使用了未分词的外部语料库来预训练字向量，并随机初始化词向量，且在后续训练时进行参数更新；后者在预训练字向量的基础上，使用了 Zhang 和 Clark 等人的模型作为基准分词模型，对大规模外部语料库进行分词后，无监督地预训练词向量使其拥有更优的初始值，并在后续模型训练时保持该值不做更新，以防止过拟合现象的产生。最终的实验表示这两者都达到了最先进的分词水平，这说明尽管模型结构相似而特征选取构造和预处理的实现思路不同，但是通过加入不同级别的语义信息来增强输入特征确实使模型有更好的表示能力。

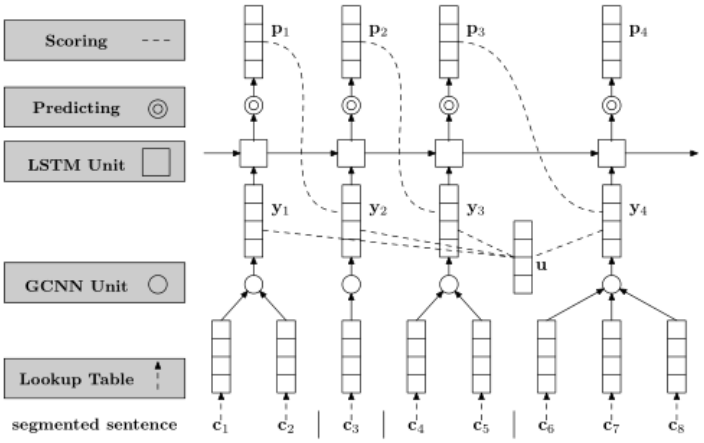

图 2-2 Cai 等人提出的神经网络分词模型结构图

2.5.3 神经网络词级别分词方法实验结果

既然基于神经网络的词级别分词模型相对标注模型具有一系列优势，那么具体应用于实际分词中的效果究竟几何，只有通过大量实验来测试。我们主要针对 Cai 等人 2016 年提出的神经网络分词模型[12]，在其于网上公布的开源代码的基础上设计实验，并取 SIGHAN Bakeoff-2005 数据集中的两个学界常用的分词语料：PKU 数据集和 MSR 数据集进行具体实验。超参数均按照其论文中公布的，除预训练词向量外，不使用任何外部语料库和工具得出最佳 F_1 值的模型对应的超参数进行设置，并选取 PKU 中的训练集对模型进行训练后，分别用 PKU 中的测试集和 MSR 的测试集进行分词测试；还设计了用 PKU 和 MSR 训练集的合集训练模型，并用 PKU 测试集进行测试的实验。用官方的评测脚本计算模型输出相比于对应的测试集的准确率、召回率和 F_1 值并记录。我们选取以 F_1 值为评价指标，实验结果如图 2-3 所示。

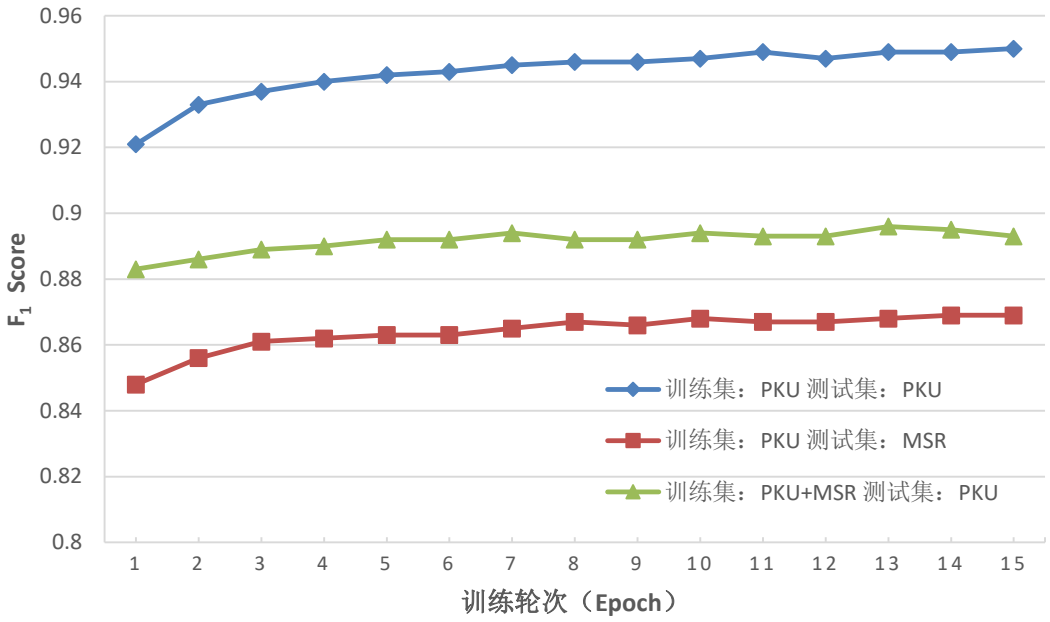

图 2-3 Cai 等人提出的模型在 Bakeoff-2005 PKU 与 MSR 数据集上的实验结果

从图中可以看出，对 PKU 训练集进行学习的模型在对应的测试集上的分词效果是最好的。而且随着对语料迭代次数的增加，F_1 值稳定地在提升，迭代到 15 次时这一数值达到了 0.95，而后续的实验显示迭代到 35 次时，F_1 值可以达到 0.953。论文中的结果表明遍历训练集 50 遍时，模型达到最优，F_1 值为 0.955，本次复现实验佐证了这一结果理论上是可以达到的，复杂的循环神经网络确实拥有很强的学习能力和自动提取特征的

能力，能很好对当前数据集上的分词模式进行建模。但是从处理分词的效率来看，复杂循环结构的表现并不理想。在我们的实验环境下（单张 NVIDIA Titan X (pascal)），训练过程中在 PKU 训练语料上迭代一次耗时一至两个小时，而迭代 35 次就需要两天以上，训练好的模型在 PKU 测试集上预测一次也需要花费数分钟的时间，这与后面列出的主流分词工具的分词速度相比大了好几个数量级，因此在实际分词应用中的意义并不大。而且从训练好的模型在不同来源的分词标注语料的表现来看，其余的分词实验也证明了不同数据集之间分词的判定标准并不统一，导致 15 次遍历 PKU 训练集集语料后的模型在 MSR 测试集上的 F_1 值只达到了 0.869，这与论文中公布的对应 MSR 数据集 0.965 的 F_1 值相去甚远。这也正符合开题报告中提到的，分词领域确实存在的一个难题：对于汉语中的词并没有一个清晰、统一、严格的定义。即使是同一届会议上由专业人员切分的内容相似的新闻语料库在分词模式上也会有所不同，这也正是同样的分词模型在这两个测试集上的准确率会有如此大的差异的原因。同时，我们也注意到即使将 PKU 和 MSR 训练集共同作为语料遍历 15 次训练后的模型，再 PKU 测试集上的 F_1 值依然保持在 0.895 左右，而且上下浮动明显。这说明两种不同分词语料的交替出现已经对于建模过程产生了较大的干扰，模型参数难以趋于收敛，神经网络虽表示能力强但也不能有效地同时对两种分词方式进行学习。这也一定程度上印证了想要直接结合多种不同分词数据集训练出一个在各个测试集上分词效果都达到最优的通用模型是不太现实的，这对迁移学习在中文分词上的应用提出了更高的要求。

此外，笔者还对四种当下主流的工业级分词软件：Thulac，Jieba，LTP，NLPIR 在 PKU 和 MSR 测试集上的表现进行了对比实验，结果如表 2-1。

表 2-1 主流分词工具在 PKU 和 MSR 测试集上的测评结果

分词系统	北京大学语料 – PKU (510 KB 大小)				微软亚研院语料 - MSR (560 KB 大小)			
	精确率	召回率	F_1 值	耗时（秒）	精确率	召回率	F_1 值	耗时（秒）
Thulac	94.1	90.8	92.4	0.154	93.3	92.5	92.9	0.176
Jieba	85.0	78.4	81.6	0.171	81.4	80.9	81.2	0.190
LTP	96.0	94.6	95.3	0.195	86.8	89.9	88.3	0.236
NLPIR	93.9	94.4	94.2	0.389	86.8	91.4	89.0	0.392

可以观察到在 PKU 测试集上分词效果最好的 LTP 和 NLPIR 工具在 MSR 测试集上都表现一般，而 Thulac 在两个测试集上都有不错的表现但没有达到的最优的水平。即使是使用了大量特征工程，加入了大规模通用词库的分词软件也难以兼顾准确率和通用性，这也一定程度上成为将中文分词作为前置工作的其它自然语言处理任务进一步提升的瓶颈。因此，解决这一问题的途径之一，正如笔者在开题报告中所提到的不应该将分词和具体任务完全割裂开来，应该使用具体任务的目标去优化分词结果，让分词服务于具体任务。既然词之间没有统一、明确的界限，那我们在实际任务中也就不需要显式地输出分词结果。直接将字序列输入多层的神经网络中，发挥其自动提取特征的能力，从字序列中获取有效的词级别的信息，并提供给下一层网络作为实际任务要建模的特征。虽然思路很简单直观，但由于缺乏可借鉴的经验，实现起来还是十分棘手，这也是未来中文分词研究工作着重要克服的难题。

2.6 现有神经网络分词模型的总结和反思

基于词序列搜索的分词模型的天然优势在于，可直接挖掘词汇本身的特征信息，甚至在进行最优分词结果搜索的同时间接获取一些句法甚至语义的信息，这等同于在分词的同时进行句法分析的任务，从而将“分词”和“理解”两个以往相互独立的过程有机地结合起来。如何在此类方法的基础上研究出一种不同于以往分词模式的高效且准确的分词方法，并验证该方法对实际自然语言处理处理任务性能的提升，减少分词任务对搜集并构建分词词典的依赖，使其具有广泛的可扩展性，这是一个值得思考的问题。

除了将深度学习应用于中文分词的研究，也有一些学者关注于不依赖于分词预处理的结果，而直接利用字级别的特征对上层自然语言处理任务建模。例如从增强输入特征表示的角度入手，探索如何利用不同粒度的语义信息，如 Johnson 和 Zhang[17]于 2017 年提出的基于深度金字塔神经网络（DPCNN）的文本分类模型。如他们在原本提出的“区域（Region）”向量[15]的基础上，依照 He 等人 2015 年提出的深度残差网络的结构[24][25]，在加深网络深度的同时，保持特征图的大小，并在每一个池化层进行下采样，使下一层的运算量减少一半，从而使得总运算量维持在原始输入文本长度的同一级别，并在多个情感和主题分类数据集上获得了准确率的显著提升。他们于 2015 年提出的区域向量表示的概念类似于卷积神经网络（CNN）中的局部感受野，由文本中当前词与若干个最邻近的词的独热向量组合而成，区域大小即所包含的词数相当于 CNN 中卷积核的大小，词向量的维度相当于输入图像中每个像素单元的通道数，唯一不同是像素是以

二维的形式排列，而词是以一维的形式。作者认为这样构造的区域向量相比于一般的词袋方式可以保留文本中的词序信息，因此区域向量应由相邻的独热向量前后拼接而成。但是 DPCNN 最终选择了以词袋方式组合的区域向量作为模型的输入，这是因为模型本身是由深度卷积网络构成，网络本身就有保留局部词序信息并逐步提取长距离全局信息的能力，因而输入词的窗口是否保留顺序就显得没有那么重要。文本中的区域向量相比于图像的一个特点在于可以根据分布假说对其进行预训练。除了独热向量作为输入外，作者也讨论了加入无监督地在原始训练集上预训练实值的区域向量的影响[16]（训练方式与 Word2Vec 类似，以最大化邻近区域对于输入区域的条件概率作为训练目标函数），发现这一额外特征对模型分类效果的有极大的提升作用，甚至单独输入预训练区域向量的结果要远好于单独输入离散独热向量的结果，这一定程度上说明复杂的网络结构不一定比有效而丰富的特征信息更重要。

Johnson 和 Zhang 二人的工作致力于在比词高一级别的“区域”中提取特征，而同期 Zhang 等人[27]以及 Conneau 等人[26]的研究则是沿着截然相反的方向。他们忽略了诸如英语等拉丁语言中有类似空格等天然词边界的信息，直接从字母的级别提取特征，并用深度卷积网络对文本分类任务建模。这种分类模型看似对汉语等无词之间分隔符的语言来说具有更好的普适性，可以免除分词等预处理操作的麻烦。但从实验结果来看，这些模型的文本分类效果还没有到达领先水平，甚至不及之前提到的基于区域向量的卷积文本分类网络。

某种意义上，字级别的文本分类方法的表现证明了词中蕴含的语义信息比字或者更低层次的语言单位中的信息更直接，更有效，也说明分词的确是中文自然语言处理的其他任务中必不可少的前置工作。但是这些直接基于字级别的英语文本分类模型正好提供了一些借鉴思路：是否可以从无空格标注词边界的字母序列中，通过神经网络动态地提取出一些长度可变但对分类文本有帮助的字母子序列，这些子序列可以是词也可以是“区域”的级别，在此基础上利用该级别的信息进行文本分类。基于词的神经网络分词模型是近几年分词研究的新兴热点，虽然其分词精确度已经能与最好的统计学习方法相媲美，但分词速度还有很大的提升空间。不过其自动学习特征，不依赖于大量特征工程的特点使其具有广阔的发展前景。

总的来看，目前使用神经网络提升分词性能的模型的创新之处主要集中在以下几点：首先，利用合适的分布表示技术将输入文本序列中的字转化为向量表示。字的向量化表示是分词工作的前提，向量表示技术的选择显然关系到后续任务的表现，合适的字向量

将对提升分词任务乃至实际任务的性能有极大的帮助；然后，构建分词模型的网络，选择合适的网络单元将任意数量的字表示转化为词表示，以获取词级别的信息。同时，网络应能将词和词之间转移的信息，即已分词句子的语义和句法信息提取出来。一般认为采用 LSTM 能对全局的分词历史建模，提取出有用的信息；最后选择合适的目标函数，合适的目标函数可以获得性能更优的模型，例如使用最大间隔准则构建的目标函数。

2.7 本章小结

本章首先介绍了几种用于提取字或词特征向量的文本表示技术。代表了中文分词主流和前沿的两种方法：基于字标注和基于词最优解搜索的方法方法都依赖于这些表示技术。其中，基于神经网络的向量表示近年来受到了业界的广泛关注和使用，不仅因其无需标注数据，设置无需人工干预或提取特征，便可以无监督地从大规模无标注语料中自动学习得到字符或词语的低维、稠密且连续向量表示，大大减少了对专家设计特征等特征工程的依赖；而且神经网络可以对更复杂的上下文进行建模，在字或词的向量表示中包含更丰富的语义信息。

近年来基于神经网络的分词方法发展迅猛，方兴未艾，主要是因为神经网络相较于传统方法有自动提取特征的优势，减少了人工特征工程的工作，但同时能达到与传统方法相持平的准确率。而且笔者总结近几年基于神经网络的分词研究工作有以下三个发展趋势：一是大部分模型都采用了循环神经网络及其变体来对输入向量序列中长距离的上下文信息进行建模，特别是基于字序列标注的分词模型。同时，这也适用于其它的序列标注任务，包括词性标注和命名实体识别。二是基于词的分词方法由于可以利用词级别特征的天然优势，在多个标准分词数据集上有超越基于字标注方法的表现，甚至与以往拥有最高准确率的模型媲美，因此近期得到了越来越多研究学者的关注和研究。三是，无论是基于字的方法还是基于词的方法，都倾向于发掘不同层次、不同级别的语义信息。这些丰富而互不重叠的特征往往对模型的性能有很强的增益作用。

但从另一个角度来看，近年来兴起的神经网络分词模型绝大多数都依赖于复杂的循环结构来自动提取字或词级别的特征以及关联特征。而且大多还需要一个标签转移概率矩阵直接来对每个字的标注依赖信息建模，尽管可以使用一些动态规划算法，但是解码过程中所需的搜索句子级别最优标注序列的时间复杂度是相当大的。再加上循环结构的本身特性导致其只能顺序地处理序列输入，可并行性相当差，因而导致这些神经分词方法效率较低。对处于自然语言处理任务底层的分词任务来说，这样的时间复杂度是几乎

不能容忍的。因此神经分词模型虽然精度高，但在实际应用中鲜少见其身影。如何在不损失分词精确度和召回率的情况下，大大提升其处理效率，或者说如何在精度和速度之间找到一个很好地平衡，我们将在下一章节中，提出我们设计的基于纯注意力机制的分词模型，尝试在一定程度上缓解这一难以调和的矛盾。

第 3 章 基于纯注意力机制的分词模型

近两年来，注意力模型在诸如序列翻译这类序列转导模型中大放异彩，那么在对中文分词这种比较基础，不需要大量依赖于语义理解和推理任务是否也有可为之处？因此笔者借鉴谷歌团队去年提出 Transformer 模型中提出的多头、自我注意力模型，改进了其中的网络结构以及处理逻辑，使其适应于中文分词任务，设计了一种基于纯注意力机制的神经网络分词模型。下面就对原始的注意力模型和我们提出的纯注意力分词模型作详细介绍。

3.1 编码-解码模型以及注意力机制

在过去的两三年，注意力机制从开始出现，到进入众多深度学习研究人员的视野，并逐渐受到越来越多的关注，发展态势十分迅猛。注意力机制是对传统编码-解码模型的一种改进和延伸，而所谓编码-解码模型则又是为了解决所谓序列到序列（seq2seq）问题所提出的。这类模型遵循先“编码”，后“解码”的处理顺序，所谓“编码”就是将离散的符号表示的输入序列转化成一个固定长度的连续的向量表示的过程；而“解码”的过程，就是将编码器生成的固定长度向量再转化成输出符号序列。

而对具体任务建模时，编码器和解码器的结构可以相适应的进行设计，在深度学习中可以选择卷积神经网络、全卷积神经网络和循环神经网络及其变体实现编码和解码的功能。例如，Facebook 的团队在 2017 年提出了基于卷积的序列到序列（简称 ConvS2S）模型[48]，将编码器、解码器甚至是记忆单元全部替换成卷积结构，一定程度上避免了采用循环神经网络进行序列建模造成的并行处理效率低下问题。虽然单层卷积核只能看到固定范围的上下文，但是将多个卷积层堆叠起来就可以将有效的上下文范围放大，从而很容易提取全局的结构信息，因此能很好地地解决序列到序列的问题。

传统的编码-解码模型虽然能很好地对序列问题建模，但是局限性也非常大。局限之一就在于位于后端的解码器能够利用输入序列的唯一信息，也就是编码器和解码器间的唯一联系，便是位于前端的编码器输出的一个固定长度的语义向量。也就是说，在编码过程中，要将整个序列的特征信息全部压缩和整合到这个固定的向量中。这样一来便损失了大量的序列本身的特征以及前后顺序信息。并且输入序列越长，通过固定向量传递信息造成的损失也就越严重。既然一开始在解码的时候就没有获得输入序列足够的特征，那么解码的准确度自然也难以得到有效的提升。

针对编码-解码模型的这一缺陷，在 2014 年，Bahdanau 等人[49]在研究机器翻译问题时采用了一种序列对齐和转导模型，后来此模型被称为注意力模型（Attention Model）。简单的说，这一模型的主要实现思路是在编码器在产生输出的时候，并不只是产生固定长度的向量表示，还会生成一个“注意力范围”来指导解码器要重点关注输入序列中的哪些部分，然后解码器根据这些关注的区域及其重要性来产生下一个输出，如此循环往复。而注意力机制实现的函数可以描述为将一组检索（query）和一组键值对（key-value）映射为输出序列，其中检索、键、值和输出都是向量表示。一般来说，输出为值序列加权后的总和，而分配给其中每个值的权重通过定义在检索与相应键上的兼容函数来计算。

相比于传统的编码-解码模型，注意力模型最大的区别在于编码器需要将输入序列中的每个时序一一对应地转化为一个向量，从而输出一个编码后的向量序列。而轮到解码器处理的时候，再对序列中的每一个向量赋予一定的权重再进行进一步处理。这样一来，解码器在产生每一个时序上的输出时，都能充分而有效地利用好输入序列携带的信息和特征。而且与一般的循环神经网络或卷积神经网络相比，注意力模型更适合处理序列中的长距离依赖关系，解码器中任意位置的输出无需通过多次循环的方式获得任意位置的输入，也无需靠堆叠的卷积层才能利用输入序列的全局信息，而是一次性直接关注整个输入序列的所有位置。在机器翻译任务中，注意力模型能直接挖掘句子内部每一对单词甚至短语间的语义组合和依赖关系，使得翻译时更好地利用单词组合或短语甚至更高级别的信息，从而使得解码器从目标语言词表中更好地挑选出语义上相匹配的单词，因此这种方法在翻译任务中取得了相当不俗的成绩。

在各种处理序列和转导建模的任务中，注意力机制已经成为其中不可或缺的一部分，主要因为其可以无视输入或输出序列中的距离长度而直接对依赖关系建模。但是过去的一段时间里，在序列建模和转导问题，如语言模型和机器翻译中，循环神经网络特别是长短期记忆和门控循环神经网络，被公认为是最有效、最先进的方法，而注意力仅仅是连接编码器和解码器的一种机制，因此注意力模型常与循环神经网络同时出现。由于循环网络结构对上下文信息建模的典型情况，是根据前一步的隐藏状态和当前步骤的输入位置计算新的隐藏状态，从而得到状态序列。但是这种固有的顺序计算特性妨碍了模型训练的并行化，而且序列长度越大时，这一制约因素也就体现得越明显。而有限的内存又限制样本批次大小，从而成为了模型并行计算的另一个上限。尽管一些深入研究条件计算和因子分解的工作能在不损失模型性能的前提下，大大提高循环网络结构的计算效

率。然而，制约了将这一固有的顺序计算结构实现更高效的并行化处理的缺陷，还是不能得到根本上的改善。

于是一些研究工作开始尝试使用卷积神经网络代替循环结构作为编码-解码模型的基本模块，并行地计算所有输入和输出位置的隐藏表示。例如 2016 年采用了空洞卷积核的 ByteNet[50]模型以及前文提到的 ConvS2S，两者在提高模型训练速度的同时取得了较好的序列转导效果。但在这些模型中，提取任意两个输入和输出位置间的依赖关系所需的操作次数，会随着位置间距的增大而增加。对于前者来说，长期依赖关系计算量关于距离，呈对数式增长的，而后者中，这一关系则是成线性增加。一定程度上来说，这一问题使得学习距离较长的位置之间的依赖关系变得更加困难。

不同于以往编码-解码模型大量依赖于循环或卷积网络结构，在 2017 年 6 月，来自谷歌的团队在论文《Attention Is All You Need》[46]中提出了一个完全基于注意力机制的编码-解码模型，论文中称之为 Transformer。它完全抛弃了之前其它基于编码-解码框架的序列模型或转导模型引入注意力机制后，仍然使用了一部分循环或卷积神经网络结构的做法，仅仅只基于注意力机制构建网络结构。模型在任务表现、并行能力和训练速度等方面都有大幅度提高。从此该模型也成为了机器翻译等许多自然语言理解任务中的重要基准模型。

Transformer 模型的主要创新在于其颠覆了以往使用注意力机制的固有模式，提出了仅依赖注意力的神经网络模型，这离不开自我注意力或内部注意力以及多头（Multi-Head）注意力概念的提出，其结构如图 3-1 所示。其中自我注意力使得注意力机制不仅可以在编码器和解码器之间实现，也可以在编码层或解码层内部实现对序列各个位置的处理修饰，寻找序列内部的联系和依赖关系；而多头注意力则是将检索和键值对通过几组不同位置间共享的线性变换矩阵映射后，在进行注意力函数的计算，相当于把这个注意力计算过程重复多次，最后把结果拼接起来。实验结果显示几组不同的线性变换，或者说不同的注意力变换头函数之间学习到了执行不同的序列特征权重，可以在一定程度上对句子的句法和语义结构行为建模。此外，注意力机制中的兼容函数采用的是缩放版的的点积，亦即每个值的权重等于当前检索和每个键的点积，并和开根号的键向量维度相除，再使用一个 softmax 函数得到。而除以键向量维度开根号的因子则被认为能避免梯度爆炸的问题。

最后，由于 Transformer 模型并不包含循环和卷积结构，因此难以直接对序列中的前后位置信息建模。为了让模型捕捉序列的顺序，需在模型中注入一些序列中关于相对

或者绝对位置的信息，于是编码器和解码器的输入层加入了位置嵌入向量。具体实现上，将每个位置编号，然后每个编号对应一个由周期与编号成正比的正弦或余弦函数在对应维度上取值的向量，最后将位置嵌入和词嵌入加和作为输入，这样一来给每个词都引入了一定的相对或绝对位置信息，使得模型具有了对句法、语义等结构信息建模的能力。

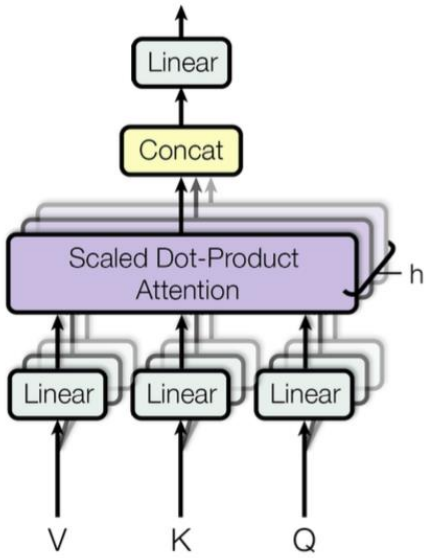

图 3-1 谷歌团队提出的 Transformer 模型中使用的多头、自我注意力单元示意图

此外模型中还加入了一些残差连接、普通前馈神经网络以及层标准化（Layer Normalization）等结构，也使用了例如 dropout 和标签平滑的技巧，最终使得模型在采用的大规模语料集的英德翻译和英法翻译任务中都刷新了当时的得分记录，且模型训练时间大大缩短。较以往成绩最优模型，该模型只需数天便可得到训练迭代收敛后的最优成绩。

3.2 单层注意力分词模型结构

在实际任务中，并非所有问题都需要获取长程的、全局的依赖信息，也有相当一部分任务只需对局部结构就能取得足够好的效果。因此论文中也提到了一种改进版的受限注意力模型，即假设当前词只与所在位置前后若干个词发生联系，相当于加入了上下文窗口的概念，使得输出序列的某一位置只能关注相邻几个位置的信息，这样一来便容易能捕捉到序列的局部结构，而这一点与前文提到的 Word2Vec 词向量建模，特别是 Skip-gram 模型具有相似之处。

词向量是由对语言模型建模得到，具体来说其建模目标是最大化样本中所有词对上下文的条件似然之积。既然是通过上下文预测当前词，那么与机器翻译根据当前语言的词序列预测目标语言的词序列类似，网络对每个位置进行推断时都需要输出词表中每个词的条件似然，所以说机器翻译与语言模型的建模目标有共同之处。Word2Vec 词向量模型有时也被认为是一种自编码器（Auto-Encoder），对于两层的 Word2Vec 模型来说，其编码器相当于是输入的嵌入层，从字符表示获得对应的编码后的连续向量。而解码器则是输出层的权重矩阵，而由于输出层的神经元数量等于词表的大小，因此该矩阵也可以视作另一个嵌入层，我们也可以从中获得任意一个词的向量表示。词向量模型的一次前向传播过程相当于输入向量通过和输出层的任意一个词向量做内积后，经过 softmax 激活函数获得输出该词的条件概率。与注意力函数的实现做类比，输入层词向量为检索（Q），输出层词向量为键（K），做内积后的值便是兼容函数的输出，而经过 softmax 激活后便得到值（V）的权重，再将值向量对这些权重进行加权就和便得到了完整的注意力函数实现。而 Skip-gram 模型一般表述为对当前词对上下文窗口中任意一个词的条件概率建模。但实际上，由于训练样本批次的存在，目标函数需要对当前词对上下文窗口中的所有词的条件似然进行加和。从某种角度上说，便等同于一次性同时关注上下文窗口中的词并赋予权重，所以说和受限制的注意力模型具有异曲同工之妙。

在笔者设计的单层注意力分词模型中，主要借鉴了 Transformer 模型中的缩放版多头、自我注意力单元且同样引入了位置向量，用来对字序列全局的输入输出依赖关系、以及结构化的顺序信息乃至高层次的句法和语义信息建模。其基本结构如图 4-1 所示。

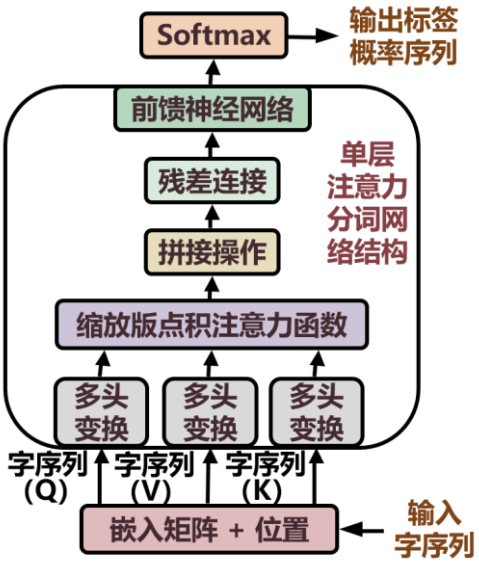

图 3-2 单层注意力分词模型结构示意图

显而易见，单层注意力模型主要由以下几部分构成，首先将字序列对应的嵌入向量序列与等长的位置向量序列相加后得到输入序列，检索（Q）和键值对（K, V）序列在这种情况下都对应于相同的输入序列；然后检索和键值对被输入几组单独的矩阵进行线性变换，也可称之为多头（Multi-Head）变换，这样便得到了多组检索和键值对。然后对每一组变换后的检索和键值对分别进行缩放版的注意力函数映射，再将每组的注意力映射输出的向量前后拼接，接着再经过一个与输入向量相连的残差连接，以及一个神经元数目比输入向量维度大一倍的前馈神经网络来提取更丰富的特征，最后通过经 Sigmoid 或 Softmax 函数激活后的线性映射得到最终输出。

3.3 多层注意力分词模型结构

将自我注意力机制应用中文分词任务，一个首先要解决的问题便是如何将在机器翻译任务中输入注意力单元的检索、键值对向量和待分词句子的字序列输入对应起来。在原始模型的编码层和隐藏层中，自我注意力机制的检索和键值对均相同，来自上一层的输出序列，从而使得该层序列中的所有位置均可关注于上一层输出序列的每一个位置。

在具体的中文分词任务中，我们也可将自我注意力单元的输入设置为字向量序列，并使得分词标注信息直接关注字序列全局特征得到。但是正如前面几个章节中提到的，若仅仅利用字级别的信息，而不对词级别的信息建模，则难以在分词效果上更进一步。于是我们尝试不止把句子字序列作为输入，而且还从句子中提取连续的多元组（n-grams）序列经过嵌入层后作为键向量输入单层注意力单元中。所谓多元组或多元字段是指长度大于一的字符片段，我们分别将语料中出现频数大于一定阈值的二元组、三元组和四元组提取出来，并加入到字符查找表中，并构造了这三种和字序列等长的多元组序列。而且将更大的多元组作为键向量输入更深一层的自我注意力网络层中，而上一层输出的 ReLU 激活值则作为检索和值，如图 3-3 所示。从而使得多层多元组注意力分词模型可直接利用词语和成语等局部的结构信息，从而更好地解决中文分词等序列标注问题。

为了使多元组序列和字序列等长，我们在句子边界加入相应数量的特殊字符进行填充，这一点和卷积层结构中的 padding 处理类似，事实上我们也可以在字序列输入后加入卷积层来对局部的多元组信息建模。但是卷积核的参数权重是共享的，由于自然语言中词语特征的稀疏性，即使我们增加卷积核的个数，提取多元组特征的效果也不一定比把每一个中高频的多元组直接映射为随模型训练更新的嵌入向量效果好。因此，我们依然是基于纯注意力机制的模型进行分词学习。

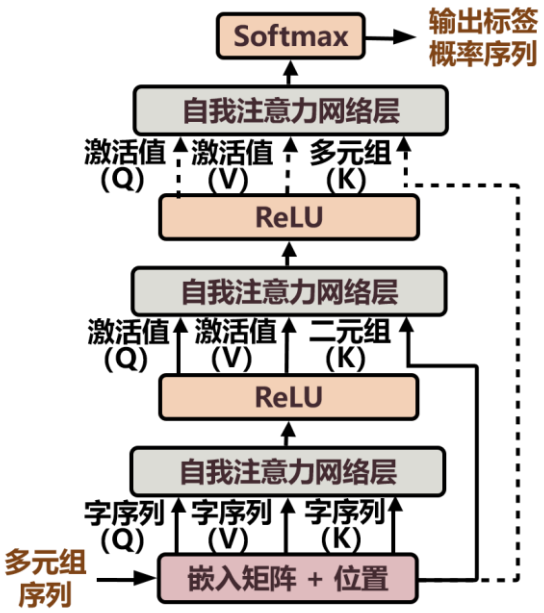

图 3-3 多层注意力分词模型结构示意图（自我注意力层即为上图中大框内部分）

3.4 本章小结

注意力机制最明显的优势在于能够一步到位、直接快速地捕捉到序列全局每个位置的联系，相比之下，循环结构需要一步步逐渐递推才能长距离依赖建模，而卷积结构则需要通过层叠网络来扩大感受野。但注意力模型的一个先天缺陷在于无法对位置信息进行很好地建模，尽管许多研究人员试图通过引入位置嵌入向量解决这一问题，但是事实上注意力机制在对序列结构化的信息比如，前后顺序关系建模能力还是较弱。而且在处理像序列标注更小、更结构化的语言理解任务，甚至更简单的算法任务，比如序列复制、倒序输出、整数加法之类的纯算法任务时，基于纯注意力机制的模型表现出来的水平并不如目前最先进的模型，甚至一般的循环神经网络。而之所以在应用于具有大规模语料的机器翻译任务上表现抢眼，是因为机器翻译任务本身并不特别强调语序，而关键在于对两种语言之间对应单词关系建模，此外翻译任务的评测指标也并不十分注重语序的正确性，即单词顺序是否完全与标准答案相符。因此，一定程度上来看，位置嵌入向量所带来的顺序信息对机器翻译任务来说已经足够了。但是对于我们提出的基于纯注意力机制的神经网络模型来说，它是否也适用于对序列结构信息要求较高的中文分词任务，甚至在速度和精度上都超越以往依赖于循环神经网络和标签转移概率的分词模型呢？这就需要进行大量的实验进行验证，在下一章节中我们设置多组在公开数据集上的实验证明我们的模型兼具高效和精度。

第 4 章 基于纯注意力机制的分词模型实验

实验采用的数据集全部来源于前文所提到的 Bakeoff 2005 官方主页上公开的中文分词数据集。该数据集分别由来自四个不同学术机构和组织的专业人员对通用语料标注而成，这四种语料来源分别是由北京大学(语料简称 PKU)、微软亚洲研究院（语料简称 MSR）、台湾中央研究院（语料简称 AS）以及香港城市大学（语料简称 CityU 或 CU）提供，其中前二者为繁体中文语料，基于 Big5 字符集编码，后二者为简体中文语料，基于 GB2312 字符集编码。但官方也提供了相应的转码成 utf-8 编码的语料以便来自不同地区的研究人员进行分析和使用。而在每种语料库上又分封闭和开放两种测试约束环境设置：基于封闭测试环境的模型只允许使用从对应的训练语料中获取的知识来进行分词学习，禁用的外部知识和材料包括且不限于词性标注信息、外部语料词频计数、阿拉伯及中文数字符号，汉语姓氏、成语等；开放测试则不受这样的约束，可以任意使用额外的语料或知识库。

4.1 分词前预处理

数据预处理的主要目的是将原始的非结构化的信号或符号等表示信息转化为计算机可处理的数据类型。而诸如数据清洗、去重、筛选等额外的预处理工作则可以减少噪声的影响，提高数据质量和纯度，模型训练效果更好，是所有自然语言处理相关的任务中都不可或缺的前置步骤。Bakeoff 的所有训练语料都遵从特定的格式规定：首先一个句子占一行；再者所有的词和标点符号都使用空格分开；最后语料中不包含词性标注信息或任何其它的标注符号，即使在机构提供的原始语料中包含了这类标注信息，官方也已提前对其剔除处理。既然所有的标点符号都应该和词分开，这也就意味着除破折号以外标点符号都应该在分词系统的输出结果中与句子的其它部分用空格分隔，那么我们也可以在训练分词模型时把每个标点符号当做句子边界处理。

而事实上，笔者在分词实验中也进行了这种预处理操作，也就是将语料按标点符号（除破折号外）分句。再从训练语料中每个句子的词序列按标签映射得到相应的字序列和标签序列，并建立相应的字符查找表（lookup table）和标签查找表，从而将字序列和标签序列转化成长度相等的两个一维整数向量，这便是训练集的一对样本。而在测试语料中则直接将标点符号分开的句子通过字符查找表表映射成一维的字符编号序列，当获得模型预测的标注序列后再按标签查找表获得每个字的标签从而得到分好词的结果，最

后将这些分好词的句子按原来的标点符号顺序组合成最终的测试集分词结果并写入文件中。

由于我们要处理的是中文分词语料，因此全部的标点符号都是全角字符，不仅如此语料中绝大部分阿拉伯数字及英文字母都是以全角字符的形式出现，占用两个标准字符（或半角字符）的位置，因此一些学者选择将这些全角字符统一转换为对应的半角符号处理。但是从一般分词习惯上来考虑，对于英文单词、阿拉伯数字甚至连续出现的中文数字通常不会切分开，因此在预处理语料时可以将其视为一个整体，所有的英文、阿拉伯数字和汉语数字片段都统一替换成对应的单字符号。从而也统一了这些字符片段对序列标注建模的影响，实验证明这一预处理方法对整体分词效果是有一定提升的，因此在以下介绍的实验中除非特别声明，都默认使用了该预处理方法。但值得注意的是，在模型输出测试集上的分词结果后，需要将这些替换掉的符号片段全部按顺序恢复到输出结果中，这样与标准分词比对时才不会产生意想之外的差错。

最后，因为深度学习框架在进行神经网络的前向传播时，是同时对同一层的所有神经元和批次中的所有样本并行做矩阵乘法，所以需要使单个批次中所有样本向量的维度相等，但在自然语言处理中，句子的字数显然是变长的，而且一般来说长度相差很大，从几个字到几百个字不等。因此，在图像领域常用的预处理方法，即把所有样本向量统一填充零值到最长的向量维度，或截断较长的向量这些做法在自然语言处理中都不太可取。如果较短的句子样本向量填充至一个固定值则会大大增加矩阵乘法的计算量，而将较长的句子截断则会损失相当一部分信息，使得模型要拟合的样本分布发生变化。于是，为了使同一批次内的输入向量等长，我们需要将语料中所有的句子按长度（由于我们要处理的是字级别的特征，在这里是字数）排序，长度相近的句子归于一个批次中，并将较短的句子向量填充零值直到和批次中最长的向量相等，最后将批次顺序打乱输入模型中进行训练，这样的预处理方法在不明显提高计算量的同时保证了模型的精度。同时为了构建验证集以比较不同超参数设置下模型的分词表现，我们将原始训练集中按长度排序的批次打乱后，取前五分之一的训练批次为验证集而后五分之四的批次为训练集。取五分之一的训练样本作为验证集，这样既不会对训练过程产生较大影响造成训练样本分布的偏差，又能使在验证集上的测试结果具有一定统计意义。当然我们也可以采取多折交叉验证的手法，以获得模型的融合结果且超参数调节的结果更具有代表性。但为了简单起见，不过度调参，我们没有采取这种方法。而且训练集的批次会在每一个训练轮次中重新洗乱，以达到更好的模型训练效果。

另外，即便填充部分为零值，在经过一些激活函数映射之后填充部分变成了非负值。为了避免填充的部分对模型推理（Inference）造成未知的影响，我们可以在模型的输入层传入一个记录每个批次中样本实际长度的整数向量，并在每次神经元被激活前或激活后应用该值，使得样本序列中对应填充部分的激活值仍为零，用以掩盖填充或未知的位置甚至其它一些需要限制的部分带来的影响。因此在神经网络中单独处理这一部分功能的结构又被称为掩藏层（Masking），这一处理手法通常也对模型的训练和预测的精度有小幅的增益作用，在随后的实验设置中我们都默认采用了此类批次划分和填充预处理及掩藏层。

4.2 注意力分词模型实验

在模型的分词实验中，除了实现基本的单层注意力单元分词模型，以及一些模型超参数调节的实验外，我们还将讨论将多元组分别作为检索、键向量或值向量输入注意力模型的意义和效果。此外，不同长度的多元组应以何种方式输入网络层，是以层叠的方式输入不同层级的注意力单元，还是以循环地方式输入同一个注意力单元，实现类似通用 Transformer 的结构[47]，我们尝试在实验中寻找对更适合中文分词的注意力模型结构。

由于注意力单元的输出随后可能需要加入残差连接，以增强网络层对输入扰动的拟合能力。为了不额外引入模型参数，在原始 Transformer 模型中将多头线性变换中每组的神经元大小设置为值向量维度除以多头变换的头数（即 h 值），这样一来前后拼接的向量维度等于输入值向量的维度，可直接在两者之间进行残差连接。但是由于拼接后的向量会经过一次线性映射，因此可将该映射输出的神经元大小设置为与输入向量维度一致，这样也可直接进行残差连接。因此，在随后的实验中，我们将验证残差连接以及调节多头变换的神经元大小对模型拟合能力的影响和作用。此外，我们还将讨论原模型中使用的 Dropout 以及层标准化（Layer Normalization）或批标准化（Batch Normalization）对防止梯度消失或爆炸、减轻过拟合现象和加快模型收敛速度以及增强泛化能力的作用。

在以下所有的分词实验中，我们使用的是官方提供的评分脚本将模型对测试集切分后的文本与对应的 Bakeoff 2005“黄金”标准分词文件进行比对，打印出分词测评结果，从而得到分词测评的各项评分。Bakeoff 提供的评分脚本是基于 Perl 语言编写的，其评分原理是计算预测的分词结果和标准分词文件中对应每一行之间的编辑距离，即插入、删除和替换字数以及完全正确分词的词数并打印出结果，最后统计出总共插入、删除和替换的总字数，以及正确的总词数，再分别除以预测分词的总词数和标准分词的总词数

得到模型的精确率和召回率及两者的调和平均，即 F1 得分（F Measure）。在大多数研究实验工作中，都取精确率。召回率和 F1 值作为对比模型分词质量优劣的标准和依据。此外，评分脚本还会根据相应训练集的词表和词频统计得到模型对未登录词的分词精确率、召回率、以及已知词的召回率等。这些数值可以进一步用来评价模型对未登录词的识别能力。

此外，在基于字序列标注的分词模型中，一般默认采用 4 标签体系（BMES）给句子中的每个字标注其在词中的位置。但事实上，我们只需要对序列中的每个字采取两种操作便可以实现任意的分词结果，即把当前字和前一个字分开，还是接在前一个字的后面。这也是基于词的分词方法或者转移方法的用来实现分词的两种操作。我们也可以在实际分词实验中仅采用“切分”和“合并”，这两种标签对序列的每个字进行标注，这是一种比 4 标签体系更简单直接的分词操作，从而使得原来模型需输出在四个标签上的条件概率分布的多分类问题，转换成只需输出其中一个标签上似然估计的 sigmoid 函数激活值的二分类问题。而且只是对每个位置的输出值进行激活，相对传统的序列标注问题简单了许多，模型进行预测也更高效，因此我们在以下实验中也对比了这两类分词标签设置对模型学习中文分词能力的影响。在以下实验中，除非特别声明，我们都选取 Bakeoff 2005 的北京大学语料集进行模型训练和分词性能测试。

4.2.1 字序列作为输入的实验

我们首先将句子的字序列经过嵌入层之后和位置向量加和作为模型的输入。这一输入同时作为多头、自我注意力单元的检索和键值对输入模型，并且如前文所述加入了 Dropout 层、残差连接、批标准化以及一层普通的前馈神经网络以增强语义特征表示能力。我们首先参照 Transformer 模型设置其中的绝大部分超参数，包括将嵌入向量和检索即键值对的向量维度都设为 512，设置 8 组不同的多头变换矩阵，采用缩放版点积加 softmax 函数作为兼容函数，注意力函数输出的 8 组向量拼接后维度仍为 512 维与输入向量大小保持一致，后接一个输出维度不变的矩阵，并使得其神经元以 0.5 的比率在训练中进行丢弃（Dropout），随后进行残差连接和批标准化，最后经过一层神经元大小为 2048 前馈网络并使用 Relu 函数激活，最后通过一层线性变换并激活输出在每个分词标签上的权重或概率值。训练模型的损失函数为交叉熵，或者说是模型输出的标签概率分布的对数在训练样本标签分布上的期望，并加入了 L2 正则化，并采用 Adam 优化器。

我们首先通过实验分析输入向量大小对分词效果带来的影响，图 4-1 为输入向量维度分别为 128 维、256 维、512 维、1024 维以及 2048 维时模型随训练轮次增加在验证集上分词结果的 F1 分数，批次大小为 128 个样本，一个训练轮次约 1300 个批次：

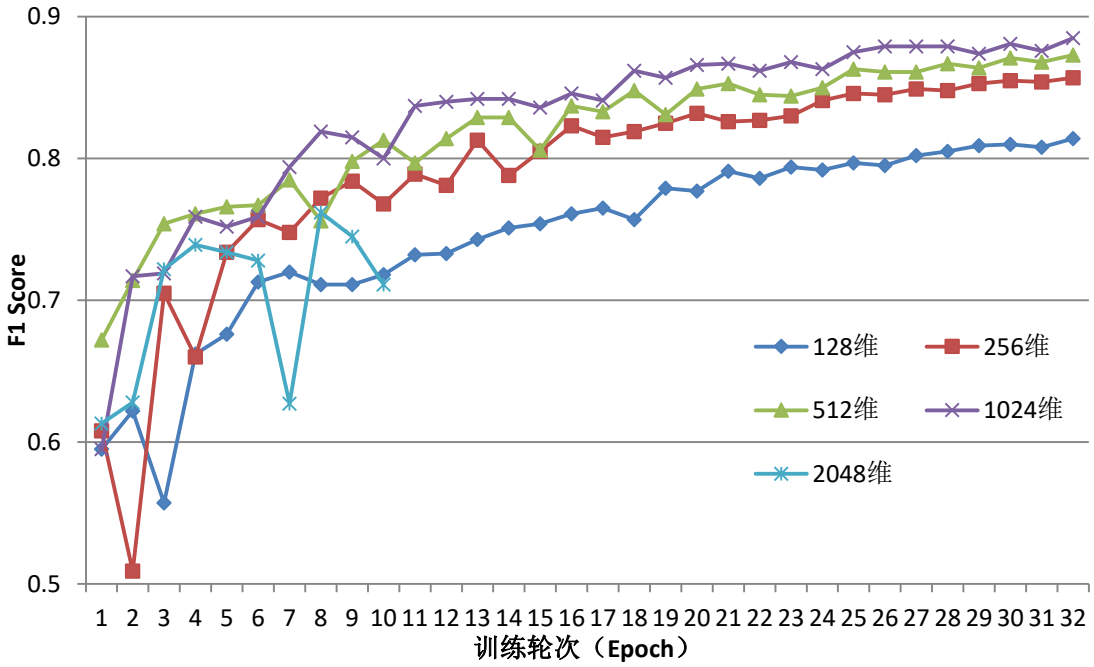

图 4-1 输入向量维度变化时随训练轮次增加在验证集上的 F1 分数

我们可以清楚地看到在但输入向量维度较小时，例如只有 128 维时，对字符的表示能力还较弱，因此模型的精度不高。而在维度变成 2048 维时模型难以进行训练，输入向量权值没有收敛，甚至可能出现了过拟合的情况。而将其设置为 512 维或者 1024 维时模型在验证集上的表现较好，而且随着训练轮次增加模型逐渐趋于收敛。注意到即使只取在训练语料中出现过一次及以上的字符加入字表，字表的总大小也不过几千维，因此模型训练时速度较快，占用内存也较小。而如果利用词级别的信息，例如语言模型、文本分类等任务中，将所有词甚至多元组加入词表的话，词表大小动辄可达数十万甚至上百万，再加上每个词向量数百维的维度大小，导致词嵌入层成为这些任务的深度学习模型中最主要的参数，对模型的最终泛化能力起到了决定性的影响。而这也体现出词法的信息，对处理一些对句法等结构建模要求不那么高的模型来说有多么重要。因此我们在之后的实验中也将引入多元字段，使得注意力模型得以利用可能成词的片段之间的长距离依赖关系。

接下来，字表中字符在训练语料中出现的最小频数对分词学习的影响，图 4-2 为最小字符频数二到五次时模型随训练轮次增加在验证集上分词结果的 F1 分数。输入向量维度大小取 512 维，其余超参数设置与模型结构与上述实验相同：

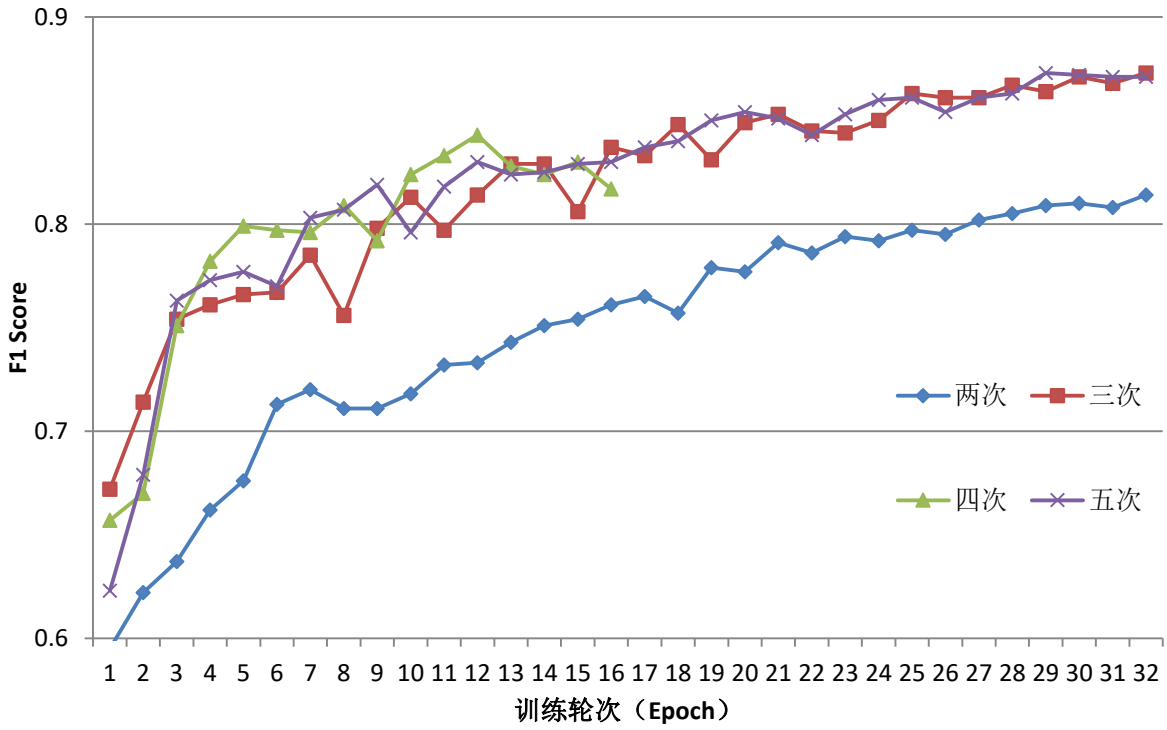

图 4-2 设定字符在语料中出现的最小频数时验证集 F1 值变化

可见在最小频数为 2 时，输入序列中引入了一些低频字造成的噪声，使得模型拟合真实分布的表现较差，而且这些低频字的嵌入向量训练不完全，可能经过几轮之后还未收敛，因此模型实际在分词效果不如将最小词频设为三到五。而另一方面，该数值也不宜设置得过大，这样的话可能使得输入序列的特征信息不够完整，难以对分词任务进行很好地学习。

然后，我们分析批次大小实际会对模型训练结果产生多大的影响。图 4-3 为最小字符频数二到五次时模型随训练轮次增加在验证集上分词结果的 F1 分数，最小字频设定为三次，而其他超参数设置与上述相同。

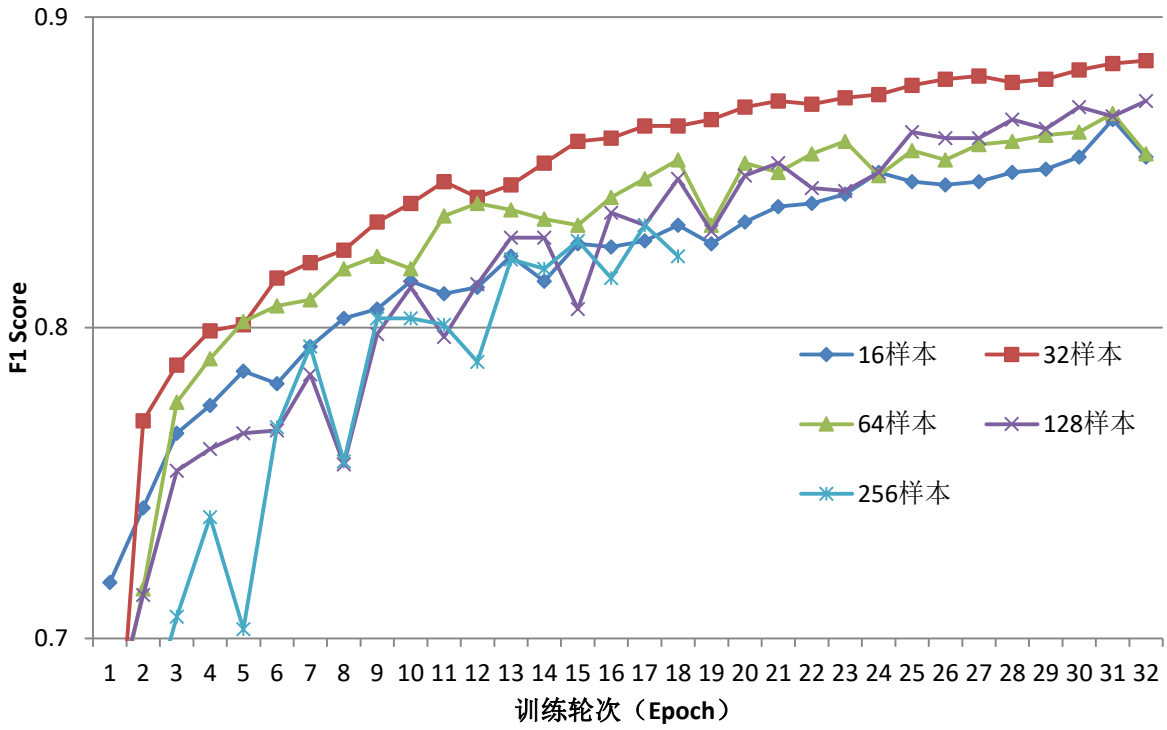

图 4-3 设定不同训练批次大小时验证集 F1 值变化

显然我们看到，一个训练批次样本的样本量越小，模型收敛的趋势便越平缓，但设置过小的值可能导致模型收敛速度较慢，训练速度也较慢。在批次很小时模型参数可能会在极小值点附近来回摆动，从而无法收敛到最优值。而设置较大的值会导致模型表现波动较大，甚至难以趋于收敛。但是设置较大的训练批次的好处在于，只要一个批次内的样本容量不会使得训练过程中内存超限溢出，那么由于矩阵乘法的并行式进行运算，而递归地求偏导数计算梯度并更新网络参数的次数会大幅减少，因此模型训练速度会得到相应提升，这一速度增益在使用图形处理单元（GPU）加速运算时表现得尤为明显。因此要在保证模型训练效果波动不大的情况下，尽量选择较大的批次。

我们还比较了在位置嵌入向量中对每个向量维度选择的正弦或余弦函数最大周期，对位置信息建模的影响，如图 4-4 所示。训练批次设定为 128，而其余超参数不变。

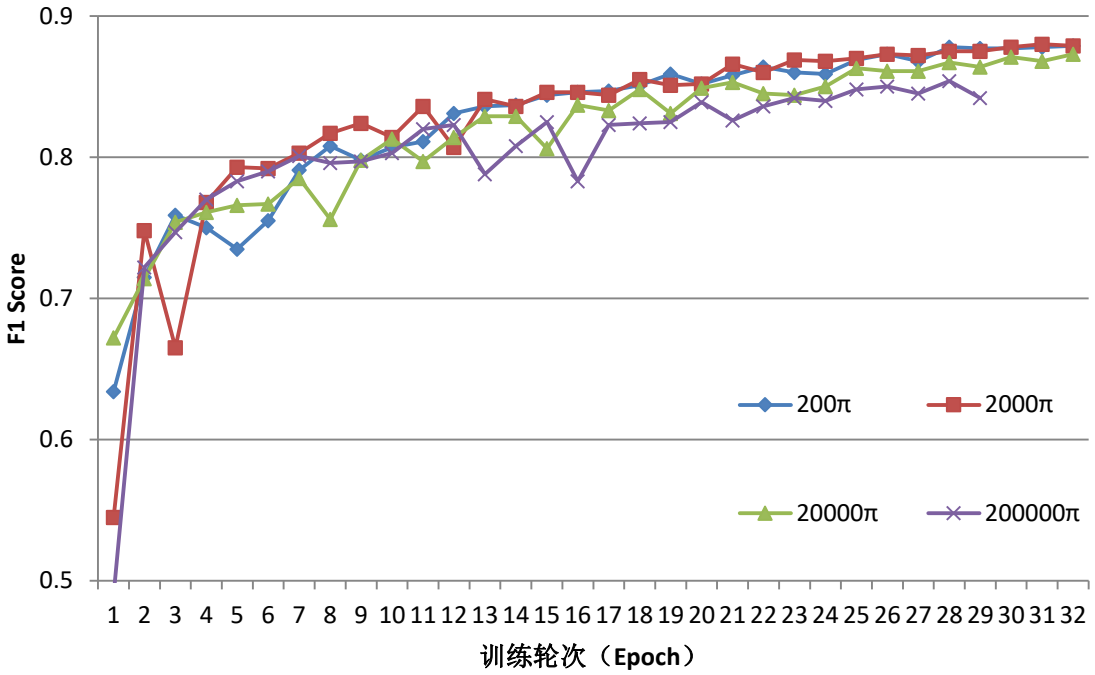

图 4-4 位置向量维度上的最大周期不同时验证集 F1 值变化

由此可见，为了使得注意力模型利用句子前后顺序这类结构信息而引入的位置嵌入中，向量维度上的三角函数最大周期的设定对位置信息的建模效果是有一定影响的。在嵌入向量总维度为 512 维，且句子长度不超过 100 个字时，三角函数随嵌入维度变化的最大周期为 2000 π 时较为适合。即三角函数的周期在 2 π 到 2000 π 之间时对字符在句子中的位置表示效果较好，过大的周期导致输入向量加和时对字嵌入表示本身产生较大的偏差。但实际在 Transformer 模型中对应的最大周期为 20000 π，这可能与任务目标不同或语言本身的的差异等一些因素导致。

同时如前文提到的，我们对采用四标签体系和两种分词操作实际达到的效果进行对比，如图 4-5 所示。位置嵌入维度的最大周期仍然设置为 20000 π，其余超参数不变，值得注意的是前面的实验都是基于两种分词操作的模型进行的。

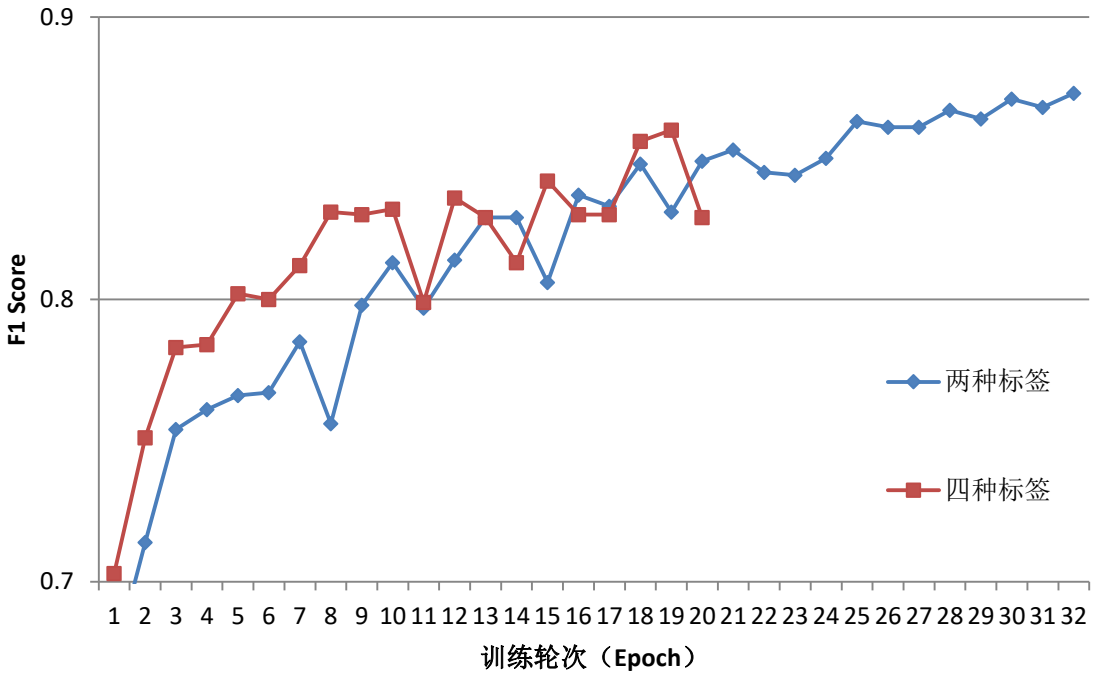

图 4-5 采用不同的分词标注体系时验证集 F1 值变化

由图表可以看出，在训练前期，采用四种分词标签体系的效果较好，但随着训练轮次的增加，四标签体系的实际效果开始明显波动，所以我们提早停止了其训练过程。而在调节其它超参数后，我们发现采用这两类不同标注模式实际上分词效果的差别很小，而在大多数情况下，采用两个分词操作的方法具有微弱的优势，我们可以将其归因于自我注意力模型本身并不具有直接对标注历史信息，或标签的前后关系及约束建模的能力，也并没引入类似 CRF 这类直接可以对标签转移概率进行控制得结构，这就导致使用较为简单地标注方法更为适合，使序列标注问题变成一个对序列中每个位置进行二分类的简化形式。而且如果要直接利用标注的顺序信息，势必要引入循环神经网络结构，例如传统的编码-解码结构，因此在接下来的实验中，我们主要采用“切分”及“合并”两种分词操作对字序列进行标注。

我们也对多头（Multi-head）注意力单元中对检索及键值对进行线性变换的矩阵组数，即头数的不同进行分析。在 Transformer 原始模型的编码器中，采用了六层的自我注意力单元，而在检索和键值对输入注意力单元后会首先经过八组不同的矩阵变换后才对其应用注意力修饰，原论文认为这几组不同的变换函数能够学习到如何处理目标不同的任务，即对句子中的词法、语义甚至句法信息有不同的侧重点。我们在实验中测试了 1 到 16 组的多头变换所带来的影响，如图 4-6 所示，其余实验设置保持与前面一致。

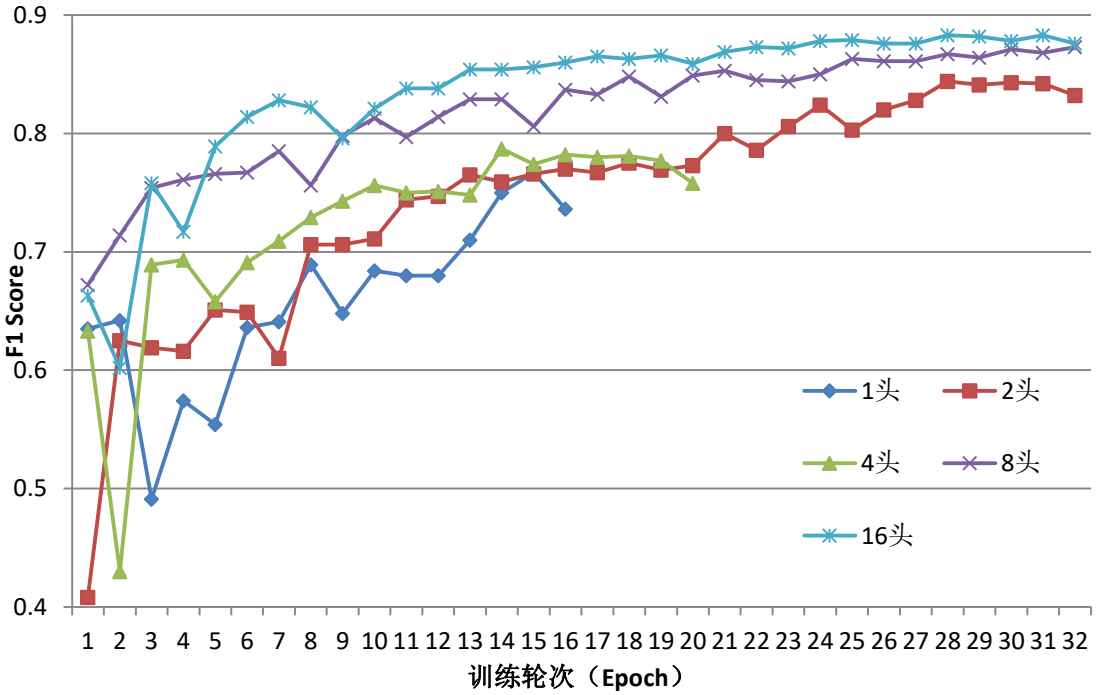

图 4-6 注意力单元的多头变换数不同时验证集 F1 值变化

从图表中我们容易观察到，头数较少的情况下模型的分词性能较差，而多头变换组数达到 16 头时，在验证集上的效果最好，这从一定程度上印证了不同的多头变换矩阵之间可能对不同的序列位置或语义等信息进行了关注，即每个多头变换的权重侧重点不同。但也并不是多头变换的组数越多越好，在原始论文中，设置 16 组多头变换的效果最佳，而在对北京大学数据集的分词实验中，但头数超过 16 时，模型的分词性能也出现了明显的下降。因为过多的多头变换会让输入注意力函数的向量维度过小，所以每个位置对序列全局的关注信息来源过于分散，这可能是导致出现这一情况的原因之一。

除了以上的实验，我们还分析了前馈神经网络的神经元维度、是否加入批标准化和残差连接，以及不同的正则化技术，例如丢弃（Dropout）和 L1 及 L2 正则化系数，对模型分词泛化能力带来的影响。实验中我们发现，前馈神经网络输出维度大小为 512 维至 2048 维时在验证集上的效果较好，前馈神经元的个数并不是越多越好，过多的参数会导致过拟合，合适的神经元数量才能较好地对语义表示信息建模。而在一般情况下，加入批标准化以及在自我注意力单元后加入残差连接，对防止出现梯度爆炸或消失、抗过拟合能力以及稳定的模型收敛都是有帮助的，但在实际实验中批标准化对分词精度产生了负面影响。而最后，合适的丢弃概率以及正则化项的系数是防止模型在学习训练样本分布时出现过拟合状况并增强模型泛化能力的关键，两者互补的看过拟合能力，针对不同的模型结构及网络层参数设置合适的丢弃概率以及正则化项十分重要。

4.2.2 字序列以及二元字段作为输入的实验

与上一小节中的实验设置不同的是，我们将从语料中提取的字序列和二元字段序列共同作为输入特征，具体来说便是将字序列作为检索和值向量序列，而二元字段则当做键向量序列输入注意力单元。此外使用的模型结构与网络层超参数仍与上一小节中模型在验证集上表现最优时的设置一致，并且仍然只使用单层的自我注意力单元。而由于二元组的加入使得字表大小大幅增加，输入的嵌入层参数也相应变多，我们调节了一部分与正则化有关的系数值，使得模型训练不至于过拟合。而单独使用字向量序列作为输入和同时使用字序列和二元字段序列作为输入，对分词学习能力的影响如图 4-7 所示。

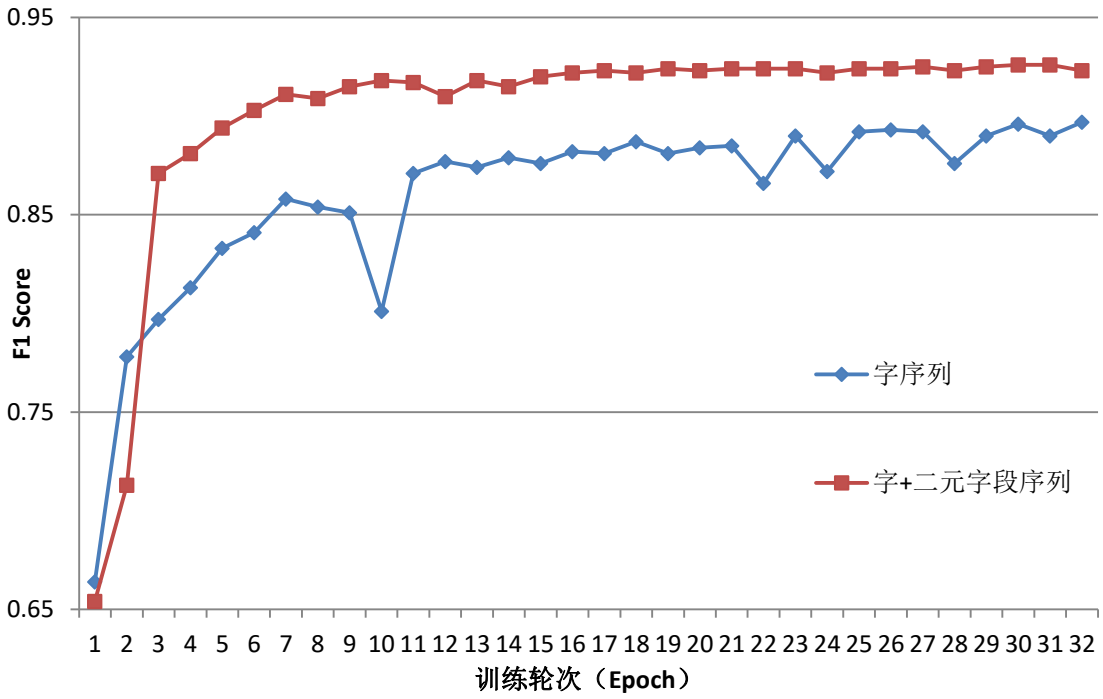

图 4-7 是否加入二元字段序列特征对验证集 F1 值的影响

可见在加入二元字段的特征后，模型的学习中文分词的能力有了大幅提升。这在一定程度上说明局部结构信息，甚至词级别的语义信息对于注意力模型对中文分词建模具有十分重要的意义。因为这种字段的加入弥补了注意力机制难以直接利用序列中的局部顺序结构特征，而只能依靠位置向量建模的缺憾。同时我们还进行了分别将二元字段作为检索或者值向量，甚至作为恒等映射输入残差连接的实验。但是最终的实验结果显示，这些输入结构并没有比二元字段作为键向量这种情况时模型分词性能更好，这也从某种角度说明额外加入的局部上下文信息更适合作为字嵌入关注内部序列的权重或依据，而不是直接被关注的特征，这样一来注意力模型所能关注的信息更多，因此泛化能力更强。

4.2.3 多元字段输入及多层注意力单元的实验

在最后的实验设计中，我们尝试同时将字序列和二元至四元字段序列作为输入特征，并堆叠了多层的多头注意力单元，使得这些输入序列分别对应不同层次的注意力函数的键向量序列，使得模型中越高层的注意力单元可以关注越大范围上下文信息。四种包含不同级别的多元组信息以及多层级的注意力分词模型在北京大学语料库的验证集上的结果如图 4-8 所示。为了避免多层网络带来潜在的过拟合作用，我们将每层的输入向量维度调整为 256 维，前馈神经元个数降为 512 个，字符或字段在训练语料中的最小频数设置为 5 次，位置向量最大周期降为 2000 π，取消了批标准化层，每层注意力单元中的丢弃（Dropout）概率调整为 0.2 并舍弃了 L2 正则化系数，其它各项超参数及设置保持不变。

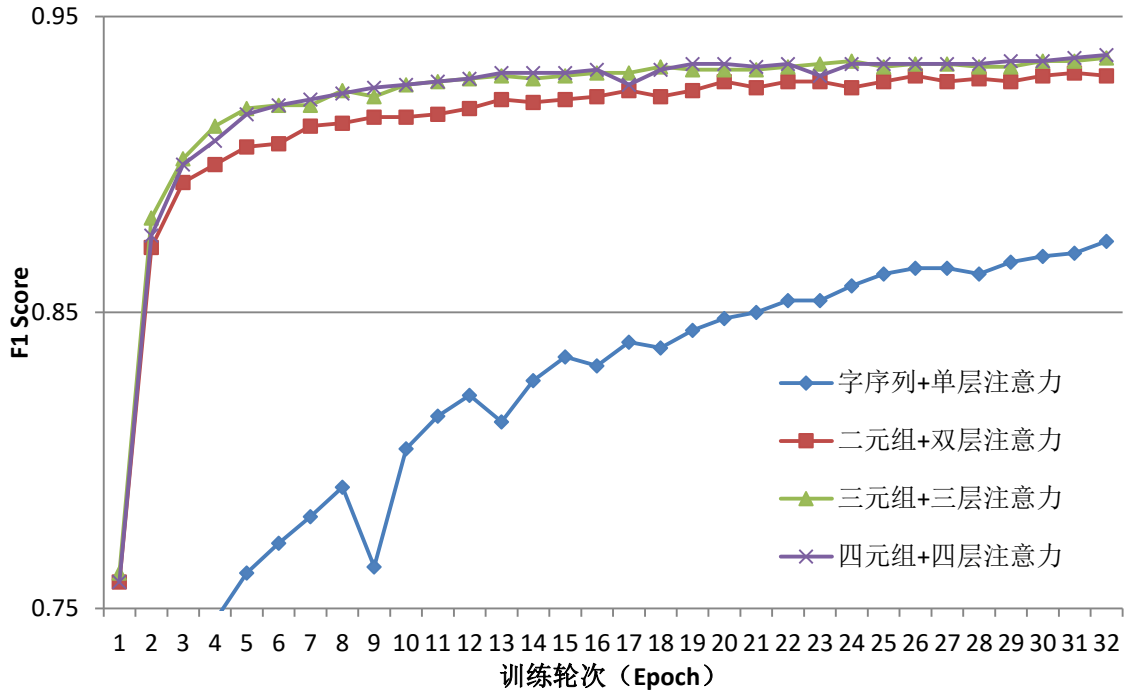

图 4-8 不同层级注意力单元以及多元组的加入对验证集 F1 值的影响

从图表中，可以看出由于嵌入向量的维度和前馈神经元的数量减少，单层注意力网络的表示能力较弱，相比于之前实验中最高 F1 分数接近 0.9 的成绩有所下滑。而两层及以上的网络加多元组的输入特征使得模型在验证集上的分词性能有大幅增益，三种模型的在训练后期分词 F1 分数均达到了 0.93 以上。此外，在将四个多元组序列特征输入四层注意力单元的在北京大学语料验证集上的 F1 分最高达到 0.938，而这一成绩已经相当接近中文分词领域在 Bakeoff 2005 的北京大学数据集的封闭测试上表现最好的模型。这在一定程度上得益于自我注意力模型对序列内部全局特征的强大提取能力，另外也归

因于多元字段的加入补全了注意力机制对于局部结构信息把握不足的潜在缺点。除此之外，我们也尝试了保持单层数量单元结构，而将字序列到多元组的序列以顺序的方式循环地输入同一层注意力单元中，从而使得一层的注意力单元能对更大范围上下文结构与语义信息建模，减少模型参数的同时，也使得网络结构更加简单。但实际上，这样的网络结构设计在分词任务中的表现并不理想，这可能是由于单层的注意力模型并不具有对复杂的较大范围上下文结构关系建模的能力，也有可能是笔者还未找到更适合这一结构的网络超参数及正则化技巧。

并且我们也注意到，两层注意力模型在只有字序列和二元字段输入的情况下表现已足够好，再加入更多的层数以及更大范围的局部上下文信息对模型泛化能力的增益并没有想象中多。这一现象在某种程度上是由于字表大小的增加导致模型网络中绝大部分权重参数都集中于字嵌入层中，增加注意力单元层数并没有对模型的矩阵参数量级产生过多影响。而另一方面，由于自然语言特征的稀疏性，使用堆叠网络层的方法获取文本序列中的层次结构信息往往效果不甚理想，反而通过预训练语言模型，从嵌入层获得字或词具有语义信息的表示向量，并在此基础上微调的效果更好。例如在文本分类任务中，一般只需通过数个词的信息便能决定一段文本的类别，而在分词任务中，通过关注上下文中几个字或词的语义或顺序信息，便对标注当前句子片段是否成词或是应该将其切分开有重要的参考价值，能很好地处理歧义或未登录词问题。因此，在大多数中文分词神经网络模型中，除了加入复杂的循环神经网络变体对上下文及标注历史信息建模，通常还要依靠大规模语料预训练的字嵌入作为输入的初始值这类迁移学习的技巧，使得模型参数收敛到更优的极值点。但这类模型由于引入了外部知识和语料库，因此属于开放测试设置规范下的分词学习结果，暂时不在本论文实验的讨论范围之内。

4.3 与其它神经网络分词模型的对比

近几年，有大量基于复杂循环结构的神经网络分词模型的研究工作涌现，这些方法通常能很好地对序列顺序结构进行建模，但是因其并行化程度不高，模型的训练和预测效率低下，在实际的工业应用中鲜见神经网络分词模型的身影。最终我们将基于字序列及二元组至四元组序列输入，拥有多层注意力单元的分词模型分别在 Bakeoff 2005 四种不同来源的分词语料（PKU、MSR、AS、CityU）上训练，并在对应的测试集上测评的结果展示在表 4-1 中。模型的所有超参数设置均与上一小节中相同。

表 4-1 自我注意力分词模型在 Bakeoff 2005 四种数据集上的测评结果

数据集/指标	精确率 P	召回率 R	F1 分数	未登录词召回率	登录词召回率
北京大学 - PKU	93.8	93.7	93.8	73.8	95.0
微软亚研院 - MSR	94.9	95.7	95.3	62.1	96.6
台湾中研院 - AS	93.3	95.0	94.2	64.3	96.4
香港城市大学 - CU	93.5	93.6	93.6	72.4	95.3

从表格容易看出，基于纯注意力机制的分词模型在四种不同来源，汉语语言用法习惯不同的语料数据集均有很强的分词学习能力及泛化性能，取得了媲美 Bakeoff 2005 封闭测评中先进模型的分词水平。与其它在相同数据集进行封闭测评的神经网络分词模型的对比如表 4-2 所示。

表 4-2 与其它神经网络分词模型在两种封闭测评的对比结果

神经分词模型	北京大学语料 - PKU			微软亚研院语料 - MSR		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
Zheng 等人[11]	92.8	92.0	92.4	92.9	93.6	93.3
Mansur 等人[55]	93.6	92.8	93.2	92.3	92.2	92.2
Pei 等人[53]	93.7	93.4	93.5	94.6	94.2	94.4
Chen 等人[52]	94.6	94.2	94.4	94.6	95.6	95.1
Chen 等人[13]	94.6	94.0	94.3	94.5	95.5	95.0
Cai 等人[12]	**95.5**	**94.9**	95.2	96.1	**96.7**	96.4
Cai 等人[23]	-	-	**95.4**	-	-	**97.1**
Ma 等人[54]	**95.5**	94.6	95.1	**96.6**	96.5	96.6
我们的模型	93.8	93.7	93.8	94.9	95.7	95.3

以上实验结果，除了 2015 年 Chen 等人的测评指标是根据 Cai 等人复现的基础上得来的，其余测评结果均来自于本人发表的论文中展示的结果。但值得注意的是，Cai 等人在模型进行训练或预测前，统一将原始数据集中的符号作半角处理，而且把英文单词和数字等不切分的都替换成固定的符号，因此其测评成绩严格意义上来说并不是公平的。但在我们的分词系统中并没有对原始数据集作任何修改，而是在代码中加入对标点、英文或数字等符号的替换和恢复，更趋近于端到端的模型，也和其它公开的测评结果具有直接可比性。这些结果在一定程度上表明，完全基于自我注意力机制，而不依赖于层叠的卷积神经网络或难以并行化的循环神经网络结构，而设计的分词模型在公开数据集上

的表现具有和先进模型竞争的能力。也说明通过一次性关注序列中所有位置全局信息的分词方法是有效的。当然我们也可以通过加入位置嵌入向量或者直接引入多元字段的方法使得注意力模型对序列顺序以及上下文语义和结构信息有进一步建模的能力。而且模型的并行化程度很高，即便在处理语料规模较大的 AS 数据集的情况下，通过单张图形计算卡（NVIDIA Titan X (pascal)）加速只需数十分钟便可完成对该注意力分词模型的训练。我们将与其它深度学习分词模型在北京大学与微软亚洲研究院语料上的训练和测试耗时对比情况列在表 4-3 中。值得注意的是这些分词速度对比都是因为这些模型的并行化程度不高，将模型训练至作者报告的最佳水平需要花费大量的时间。我们仅仅基于 Cai 等人开源的代码复现其 2016 年提出的模型时，在上述的硬件加速条件下，仍然耗费了数天才完成 50 个语料迭代轮次的训练。因为相关研究工作中使用的硬件环境和深度学习工具包都不尽相同，因此下表中的他人提出模型的分词速度对比均基于 Cai 等人 2016 年在其论文中复现他人实验的结果。

表 4-3 与其它神经网络分词模型在两种语料上的训练和测试速度对比结果

神经分词模型	北京大学语料 - PKU		微软亚研院语料 - MSR	
	训练耗时（小时）	测试耗时（秒）	训练耗时（小时）	测试耗时（秒）
Chen 等人[52]	50	105	100	120
Chen 等人[13]	58	105	117	120
Cai 等人[12]	48	95	96	105
Cai 等人[23]	3	25	6	30
Ma 等人[54]	1.5	24	3	28
我们的模型	**0.5**	**1**	**1**	**2**

从表格中的对比数据可以看出，一般的包含大量复杂循环结构的神经网络分词模型至少需要数个小时甚至数天才能完成训练，难以使用多线程以及图形加速卡实现并行运算加速，预测时的分词速度也远不如我们实现的模型。而以往的传统的分词模型又依赖于大量的特征工程，前期需要人工构造特征。因此，可以说基于自我注意力单元的分词模型在分词性能以及效率上都达到了领先的水平。

4.4 本章小结

本章从设计只依靠纯注意力机制的神经网络分词模型的思路出发，逐步介绍分词实验基于的 Bakeoff 数据集有关的测评规范，注意力分词模型网络基本结构，以及进行的

几组实验的设置还有超参数调节在相应验证集上的分词性能对比。最终和分词性能处于领先水平的其它几种神经网络分词方法作对比，证明我们设计的基于纯注意力机制的分词模型已经达到了相当优秀的水平，而且在模型训练和预测速度及并行度上，展示了突出的高效表现，因此在实际的分词场景中有非常大的应用潜力。

结论

本文从分析国内外自然语言处理问题的背景和概况以及国内二十余年来中文自动分词领域的研究进展以及一些仍悬而未决的难题出发，探讨中文分词研究在深度学习迅速崛起的背景下的机遇和挑战，阐明本论文所要探索并尝试解决中文分词性能提升瓶颈的科研方向。

在介绍了对中文分词以及其他自然语言处理任务相当重要的几种字或词的表示技术之后，我们对比了中文分词发展历史中最为实用甚至性能处于领先水平的文本切分方法。从基于字典匹配和规则的机械分词方法到能自适应不同专业领域语料的基于统计分析及文本挖掘的无词典分词法；从先分词后“理解”的传统模式，到将句法和语义信息结合在分词过程中的“分词和理解交互”的创新探索模式；从基于序列标注的分词方法再到基于词的转移操作的方法，再到近几年将序列建模能力很强的循环神经网络及其带门控制和记忆单元的变体加入分词系统的深度学习分词方法。中文分词领域的研究成果十分丰硕，研究人员努力提升中文分词性能的热情始终不减。

近两年，深度学习的研究学者将目光从卷积结构和循环结构，逐渐转向能直接并行地对序列输入中每个位置同时进行关注的注意力机制。笔者也跟随着这一脚步，探索注意力模型在中文分词任务上实现突破的可能性。依照前人在机器翻译任务上应用自我注意力机制并获得优异的翻译表现的实现思路，本文针对分词任务的特殊性，设计了一种只依赖于注意力机制的神经网络结构进行分词实验，完全抛弃了以往常在序列标注任务中加入卷积层特别是复杂的循环结构的做法。为了发挥注意力单元能一次性利用序列中全局上下文信息的先天优势，并改善其对于顺序结构以及局部信息建模能力的不足，本文在加入了位置嵌入向量的同时，设计了多层注意力单元使其能利用不同上下文范围的特征，对帮助模型进行分词任务的学习具有很好的效果。本文实现的分词模型在公开数据上表现与先进模型相当的同时并行化处理的能力较高，分词速度较快，包含数层注意力单元的分词模型只需数十分钟便可完成对大规模训练语料的学习。

据笔者目前对中文自动分词领域的了解，本文中将注意力机制应用于分词任务尚属首次尝试，并在一定程度上实现了并行化的、非顺序处理的分词系统结构在性能上的突破。这使得包括中文分词在内的一系列基于序列标注的自然语言处理任务在深度学习中有了更广泛的应用前景，可以预见在不久的将来会有更高效、并行化程度更高而且更精确的中文分词模型出现在我们的视野中。

参考文献

[1] Sproat R., Shi, C.et al.A stochastic finite-state word segmentation algorithm for Chinese [J]. Computational Linguistics, 1996, 22(3):377-404.

[2] 国家技术监督局. 中华人民共和国国家标准 GB/T13715-92 信息处理用现代汉语分词规范[S]. 北京:中国标准出版社, 1993.

[3] 孙茂松, 张磊. 人机并存，“ 质”“ 量”合一[J]. 语言文字应用, 1997, (1): 79-86.

[4] 刘开瑛. 现代汉语自动分词评测研究[J]. 语言文字应用, 1997, (1): 101-106.

[5] 孙茂松, 邹嘉彦. 汉语自动分词综述[J]. 当代语言学, 2001, 3(1), 22-32.

[6] Sproat, R.and Emerson, T. The First International Chinese Word Segmentation Bakeoff [A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo, Japan; July 11-12, 2003, 133-143.

[7] Andi Wu and Zhixin Jiang.Word segmentation in sentence analysis[A]. In: Proceedings of 1998 International Conference on Chinese Information Processing[C]. Beijing, China: 1998, 169-180.

[8] Rabiner, Lawrence R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE 77 (1989), No. 2, pp. 257–286.

[9] Lafferty, John D.; McCallum, Andrew; Pereira, Fernando C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Morgan Kaufmann Publishers, 2001, pp. 282–289.

[10] Klinger R, Tomanek K. Classical probabilistic models and conditional random fields[M]. TU, Algorithm Engineering, 2007.

[11] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C]//EMNLP. 2013: 647-657.

[12] Cai D, Zhao H. Neural word segmentation learning for Chinese[J]. arXiv preprint arXiv:1606.04300, 2016.

[13] Chen X, Qiu X, Zhu C, et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C]//EMNLP. 2015: 1197-1206.

[14] Yao Y, Huang Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation[C]//International Conference on Neural Information Processing. Springer International Publishing, 2016: 345-353.

[15] Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks[J]. arXiv preprint arXiv:1412.1058, 2014.

[16] Johnson R, Zhang T. Semi-supervised convolutional neural networks for text categorization via region embedding[C]//Advances in neural information processing systems. 2015: 919-927.

[17] Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, 1: 562-570.

[18] He H, Wu L, Yang X, et al. Dual Long Short-Term Memory Networks for Sub-Character Representation Learning[M]//Information Technology-New Generations. Springer, Cham, 2018: 421-426.

[19] Zhang M, Zhang Y, Fu G. Transition-based neural word segmentation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 421-431.

[20] Zhang Y, Clark S. Chinese segmentation with a word-based perceptron algorithm[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007: 840-847.

[21] Zhang Y, Nivre J. Transition-based dependency parsing with rich non-local features[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011: 188-193.

[22] Zhang Y, Clark S. Syntactic processing using the generalized perceptron and beam search[J]. Computational linguistics, 2011, 37(1): 105-151.

[23] Cai D, Zhao H, Zhang Z, et al. Fast and accurate neural word segmentation for chinese[J]. arXiv preprint arXiv:1704.07047, 2017.

[24] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[25] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European Conference on Computer Vision. Springer, Cham, 2016: 630-645.

[26] Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for natural language processing[J]. arXiv preprint arXiv:1606.01781, 2016.

[27] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. 2015: 649-657.

[28] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3):8-19.

[29] Salton G, Buckley C. Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24(5), 513-523[J]. Information Processing & Management, 1988, 24(5):513-523.

[30] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. Computational linguistics, 18(4):467–479, 1992.

[31] Wallach H M. Topic modeling: beyond bag-of-words[C]// International Conference. 2006:977-984.

[32] Bengio Y, Ducharme, R&#, jean, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6):1137-1155.

[33] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning[J]. Journal of Parallel & Distributed Computing, 2008:160-167.

[34] Collobert R, Weston J, Bottou, L&#, et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[35] Mnih A, Hinton G. Three new graphical models for statistical language modelling[C]// Machine Learning, Proceedings of the Twenty-Fourth International Conference. 2007:641-648.

[36] Mnih A, Hinton G. A scalable hierarchical distributed language model[C]// Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December. 2008:1081-1088.

[37] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September. 2010:1045-1048.

[38] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes[C]// Meeting of the Association for Computational Linguistics: Long Papers. 2012:873-882.

[39] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[40] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119

[41] Zhao J, Qiu X, Huang X. A Unified Model for Joint Chinese Word Segmentation and POS Tagging with Heterogeneous Annotation Corpora[C]// International Conference on Asian Language Processing. IEEE Computer Society, 2013:227-230.

[42] Jiao Z, Sun S, Sun K. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network[J]. arXiv preprint arXiv:1807.01882, 2018.

[43] Peng, Fuchun, Feng, Fangfang, McCallum, Andrew. Chinese segmentation and new word detection using conditional random fields[J]. Proceedings of Coling, 2004:562--568.

[44] 张华平, 刘群. 基于 N-最短路径方法的中文词语粗分模型[J]. 中文信息学报, 2002, 16(5):3-9.

[45] 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法[J]. 小型微型计算机系统, 2006, 27(9):1766-1771.

[46] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[47] Dehghani M, Gouws S, Vinyals O, et al. Universal transformers[J]. arXiv preprint arXiv:1807.03819, 2018.

[48] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.

[49] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[50] Kalchbrenner N, Espeholt L, Simonyan K, et al. Neural machine translation in linear time[J]. arXiv preprint arXiv:1610.10099, 2016.

[51] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[52] Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for Chinese word segmentation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1744-1753.

[53] Pei W, Ge T, Chang B. Max-margin tensor neural network for Chinese word segmentation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 293-303.

[54] Ma J, Hinrichs E. Accurate linear-time Chinese word segmentation via embedding matching[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1733-1743.

[55] Mansur M, Pei W, Chang B. Feature-based neural language model and chinese word segmentation[C]//Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013: 1271-1277.

[56] Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.

[57] John R Firth. A synopsis of linguistic theory, 1930-1955. Studies in Linguistic Analysis, 1957.

[58] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284.

[59] Brown P F, Desouza P V, Mercer R L, et al. Class-based n-gram models of natural language[J]. Computational linguistics, 1992, 18(4): 467-479.

[60] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[C]//Aistats. 2005, 5: 246-252.

[61] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 746-751.

[62] Cover T M, Thomas J A. Elements of information theory[M]. John Wiley & Sons, 2012.

[63] Baum L E, Petrie T, Soules G, et al. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains[J]. The annals of mathematical statistics, 1970, 41(1): 164-171.

[64] 来斯惟. 基于神经网络的词和文档语义向量表示方法研究[D]. 中国科学院大学, 2016.

[65] Yao Y, Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]//International Conference on Neural Information Processing. Springer, Cham, 2016: 345-353.

致谢

在北航六年半的本科及研究生的学习生活即将告一段落很荣幸，我很感谢北航这所大学，也很荣幸能加入中法这个大家庭，更为为能进入中法数据科学实验室，在于雷老师的指导下进行科研工作心怀感激。本论文没有于老师无私的帮助与教导是无法完成的。在每次科研遇到难关时都向老师请教、讨论，老师耐心讲解、悉心指导才使得我的论文进展能一直向前推动。从论文的选题到资料成果的搜集、查阅和研究分析，直至最后设计模型进行一系列实验的整个过程中，花费了于老师很多的宝贵时间和精力，获得了许多珍贵的经验和建议。在此向导师表示衷心地感谢！导师严谨的治学态度，用于创新的精神和高度的责任心都将使我受益终生。

在研究课题的过程中使我不断成长的一边是老师的悉心指导，一边是与实验室的兄弟姐妹们之间的讨论与相互学习。两年多来我的科研能力以及实验技能都得到了很好地锻炼，学习到独立推进研究工作，提出新问题，探索解决方案，整理思路设计实验并实现方案以快速验证的研究方法，也在这一过程中加强了对自我的要求，对科研甚至对自己负责的态度。我获得的不仅仅是知识、技能或是科研水平的提高还有自我的成长。

在中法数据实验室这个的团队中，我的团结意识和集体荣誉感变得更强。我也逐渐认识到知识的传承对于学术和研究的推进起到多么重要的作用。在这里，我还要感谢实验室里的几位已经毕业的学长学姐，没有他们抽空不厌其烦地回答我的基础问题，并指出我科研上的误区，使问题能即使被发现并解决，我的研究课题不可能如此顺利地起步，也不能获得如此有意义的研究成果。在此向于老师和学长学姐表示深深的谢意与崇高的敬意。