硕士学位论文

基于注意力机制卷积神经网络的社交媒体文本立场分析

RESEARCH ON KEY TECHNOLOGIES OF PARTIAL POROUS EXTERNALLY PRESSURIZED GAS BEARING

颜瑶

哈尔滨工业大学

2010 年 12 月

国内图书分类号：TM301.2　　学校代码：10213
国际图书分类号：62-5　　密级：公开

工学硕士学位论文

基于注意力机制卷积神经网络的社交媒体文本立场分析

硕士研究生：颜瑶

导　　师：徐睿峰教授

申请学位：工学硕士

学　　科：机械制造及其自动化

所在单位：机电工程学院

答辩日期：2010 年 12 月

授予学位单位：哈尔滨工业大学

Classified Index: TM301.2
U.D.C: 62-5

Dissertation for the Master's Degree in Engineering

RESEARCH ON KEY TECHNOLOGIES OF PARTIAL POROUS EXTERNALLY PRESSURIZED GAS BEARING

Candidate:	Yu Dongmei
Supervisor:	Prof. XXX
Associate Supervisor:	Prof. Assosuper
Co Supervisor:	Prof. Cosuper
Academic Degree Applied for:	Master of Engineering
Specialty:	Mechanical and Automation
Affiliation:	School of Mechatronics Engineering
Date of Defence:	December, 2010
Degree-Conferring-Institution:	Harbin Institute of Technology

摘　要

摘要是论文内容的高度概括，应具有独立性和自含性，即不阅读论文的全文，就能获得必要的信息。摘要应包括本论文的目的、主要研究内容、研究方法、创造性成果及其理论与实际意义。摘要中不宜使用公式、化学结构式、图表和非公知公用的符号和术语，不标注引用文献编号。避免将摘要写成目录式的内容介绍。

关键词：关键词 1；关键词 2；关键词 3；……；关键词 6（关键词总共 3 — 6 个，最后一个关键词后面没有标点符号）

Abstract

Externally pressurized gas bearing has been widely used in the field of aviation, semi-conductor, weave, and measurement apparatus because of its advantage of high accuracy, little friction, low heat distortion, long life-span, and no pollution. In this thesis, based on the domestic and overseas researching……

Keywords: keyword 1, keyword 2, keyword 3, ……, keyword 6 (no punctuation at the end) 英文摘要与中文摘要的内容应一致，在语法、用词上应准确无误。

目 录

第 1 章 绪论

1.1 课题来源

本课题来源于国家自然科学基金重点项目《社交媒体中文本情感语义计算理论和方法》、国家自然科学基金面上项目《文本情绪计算框架、模型和方法研究》、广东省数据科学工程技术研究中心开放课题《社会化媒体大数据群体情感深层理解与预测研究》、《深圳市孔雀计划技术创新项目结合脑科学和深度学习的文本情绪计算研究及其在社会群体情绪分析的应用》。

本章 2.2 节介绍情感分析的相关研究，2.3 节介绍立场分析相关研究，2.4 节着重介绍基于深度学习模型的立场分析研究。

1.2 课题研究的背景、目的与意义

随着互联网与移动互联网的飞速发展和多样化网络交流软件的普及使用, 越来越多的网络用户可随时随地浏览热点新闻报道，与此同时借助微博、Twiter、知乎等平台上围绕各种新闻报道、热点社会事件等话题发表自己的观点和表达自己的立场、情绪。这些海量的评论数据，对商业智能、舆情分析、政府决策等都具有重要的研究价值。例如政府想确定公民对新政策的意见和态度，可能会使用在其新政策相关的微博、论坛中的帖子来收集反馈。先阶段研究方式主要基于人工调查分析，但人工调查有成本高、时效性差等缺点。在这一时代背景下，如何利用现阶段比较成熟的自然语言处理、机器学习、深度学习等技术分析出用户的立场成为了一个待解决的问题。

文本情感分析，指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。通常来说，情感分析的目的是为了找出说话者/作者在某些话题上或者针对一个文本两极的观点的态度。这个态度或许是他或她的个人判断或是评估，也许是他当时的情感状态，或是作者有意向的情感交流。文本情感分析的一个基本步骤就是将文本中的某段已知文字的两极性进行分类，这个分类可能是在句子级、功能级。分类的作用就是判断出此文字中表述的观点是积极的、消极的、还是中性的情绪。更高级的“超出两极性”的情感分析还会寻找更复杂的情绪状态，比如“生气”、“悲伤”、“快乐”等等。

目前, 文本情感分析领域的主要方法为基于规则和基于机器学习的方法。前者

借助已有情感词典和语言学基础, 利用情感词汇单元的极性以及其他语言成分对情感词汇结合、强化、传播等作用, 达到文本的情感分类的目的。这种方法虽然无需标注大量的训练数据, 但对情感词典资源和语言学基础有比较高的要求。在大量情感新词的不断出现，很难全面的归纳出一套完整精确的情感字典，而且网络社交平台口语化，特性化的表达也给语言规定带来了巨大的挑战，因此文本情感分析的主流集中在基于机器学习的方法。基于机器学习的方法通过特征工程的方法选择特征词，然后根据特征词构造出文本的特征表示，再结合已经标注的训练样本构造情感分析模型。这一方法为文本情感分析的表现带来了显著的提升。但基于机器学习的情感分析方法却面临标示向量稀疏和特征词构造复杂等问题。近年来, 深度学习在自然语言处理领域取得广泛的进展．借助与 Word2Vec、循环神经网络等深度学习技术可以将稀疏词表示变为稠密、连续、低维的向量, 并利用端到端分类减少特征构建的复杂性。基于深度学习的情感分析方法已经取得较高的性能。

文本情感分析将文本分为正向、负向、中性等多种类型, 但是该分类方式经常无法满足很多场景的需求。近年来, 在给出特定话题或目标的前提下, 分析文本对该目标所持支持、反对、中性立场的立场分析逐渐得到关注, 并出现了一些针对网络论坛辩论、以及微博和 Twitter 语料的立场分析研究与评测。表面上立场分析与传统的情感分析任务具有较高的相似性, 实际上两者存在显著的不同。情感分析只关注文本本身表达的正向、负向的情感，无需关注其他的内容。而立场分析需要进一步分析出文本对特定的目标的立场，这种立场意图有可能是显式或者隐式的，这也极大增加的研究文本立场分析的难度。例如有关美国大选“希拉里”为主题，文本内容为“希拉里是个十足的病态的骗子”，此文本明显归为负向的情感，对于“希拉里”主题持“反对”态度，但是假设以“希拉里”的竞争对手“特朗普”为主题，文本的情感分析依旧没有改变，但是由于基于不同的立场改变，文本的立场也改变成“支持”了。在文本立场分析中, 同一个立场，其文本表达可以是正向或者负向的。文本立场分析与文本情感分析有着显著的区别，文本立场分析 1.4 本文的主要研究内容和组织结构是在文本情绪分析上的更进一步的深入研究。

本文以社交媒体文本的立场分析作为研究目的，研究基于现有深度学习框架并结合立场评价对象的信息分析社交媒体立场的方法，此方法结合了注意力机制和卷积神经网络，改善现有研究未充分挖掘立场主题和立场文本的关系，从而提高文本分析任务上的性能。该方法对文本立场分析研究有一定的科学贡献; 同时, 本文研究的立场分析方法在商业智能、舆情分析、政府决策中具有较高的应用价值。

1.3 国内外相关研究概况

文本情感分析和立场分析作为自然语言处理领域的热点问题,吸引很多的研究者的关注，研究者们也取得可观的成果。本文将从文本情感分析研究现状、文本立场分析研究现状方面对国内外相关研究展开介绍。

1.3.1 文本情感分析研究现状

文本情感分类可以按照处理粒度分为词语级、短语级、句子级、篇章级情感分类任务。目前，大部分文本情感分类研究可分为两种主要研究思路：基于情感词典/规则和知识库的方法以及基于机器学习的特征分类的方法。前者主要是依靠人工构建的情感词典或领域词典，以及主观文本中带有情感极性的组合单元，来获取情感文本的极性。后者主要是使用机器学习的方法，选取有效的分类特征构造分类器来完成分类任务。近年来，基于深度神经网络和表示学习的方法在情感分类领域得到广泛重视。因此，本节在回顾文本情感分类研究现状时，将对基于表示学习的方法进行单独讨论。

基于情感词典、规则和常识库是文本情感分类的基本方法。基本思想是在待分类文本中对情感词典记录的词语、规则和知识库条目进行匹配，利用匹配条目中记录的情感分类信息生成文本的情感分类结果。典型的工作如 [Ma et al.2005] 进行的基于词语匹配特征的方法。基于情感知识规则库的分类方法是近年来出现的方法。[Wu et al. 2006] 使用 Apriori 算法从情感标注训练语料中挖掘语义标签、属性与情感分类之间的关联规则。应用情感产生规则，建立一个分离混合模型来计算输入文本与情感关联规则之间的相似性实现情感分类。[Godbole et al. 2007] 构建了基于英文词典的情感分析系统，并用该系统来评估新闻文本中各个实体（人物、地点、时间）的情感倾向。[5] [毛峡等, 2011] 以 OCC 认知情感模型为基础，结合知网常识库进行扩展，制订了 22 种情感识别规则，对应识别 OCC 模型定义的 22 种情感类型。[6][Balahur et al.2011] 建立一个常识库用来记录各种引发情感的事件，从而识别出没有明确情感指示词汇出现的文本中的情感类别。其核心思想是把情景以本体的形式建模成一连串动作及其与其相应的情感。[7][Taboada et al.2011] 提出使用情感词典中情感词以及程度副词和否定词特征对句子或短语的情感分类。[8][Udochukwu et al.2015] 在 OCC 模型的基础上，提出了一种基于规则的隐式情感分析方法，该方法性能相较于词匹配方法有明显提升。近年来基于情感常识库的分类方法也获得一定发展。[9][Agarwal et al. 2015] 基于 ConceptNet 常识库

提取文本表达中的重要概念以及相关特征。随后利用 SenticNet、SentiWordNet 和 General Inquirer 构建上下文情感词典，依据各个特征的重要性对文本进行情感分类。[10][Tromp et al.2015] 结合 Plutchick 的情感轮模型和基于规则的方法，构建了一种社会化媒体中情感分类的框架，提高了文本情感分类准确度。总体来看，基于情感词典/规则和知识库的情感分类准确率较高，但由于情感词典和常识库规模的限制，覆盖率较低。同时此类方法对分词、词性标注、规则匹配等的准确性要求较高，系统内部错误传递影响较大。[30]

机器学习近年来基于机器学习的情感分类研究得到快速发展。[Aman et al. 2007] 应用情感词的 Unigram 特征，结合支持向量机 (Support Vector Machines, SVMs) 进行文本情感分类。[Abbasi et al. 2008] 在多种不同的数据集上选取多种不同的特征，采用集成分类器方法，实现对文本的情感分类。[11][Neviarouskaya et al. 2010] 提出情感分析模型 (Affect Analysis Model, AAM)，结合人工构建语言学规则库，解决带有语法错误、缩写、Emoticon 等非正规文文本中的情感分类。[Keshtkar and Inkpen, 2012] 提出分层分类方法实现博客作者的心情 (Mood) 分类。该方法由粗到细，将 100 多种心情类别分成五个层次。通过分层训练 SVM 分类器，先对大类进行分类，而后再进行下一级小类的分类。[12][Mohammad et al. 2012] 将情感词典与 n-gram 特征结合，使用 SVM 分类器进行情感分类。随着机器学习技术的发展，研究者提出了多种分类算法和特征集合的方法。[Xia et al. 2011] 使用词性、词的关系两大类特征集，朴素贝叶斯、最大熵和 SVM 等三种基分类器，固定组合、权重组合和 meta-classifier 三种分类器集成策略，通过特征、分类器、集成策略组合，有效提升文本情感分类性能。针对文本中跨句子的复杂语言结构，[13][Yang et al. 2014] 提出了一种融合局部和全局层次的情感分类方法。在学习过程中把语言信息编码为软约束，并通过后正规化的方法将软约束转化为条件随机场 (Conditional Random Fields, CRFs) 的特征参数，应用条件随机场解决情感分析问题。[29][张林等,2014] 提出了一种基于短评论特征共现的特征筛选方法，将短小评论中的优势信息和传统的特征筛选方法相结合，在筛选掉无用噪音的同时增补有利于分类的有效特征，有效提高短文本的情感分类效果。[张志琳等, 2015] 提出选取词汇化主题特征、情感词内容特征和概率化的情感词倾向性特征用于微博情感分类。此外，基于跨领域、跨语言迁移学习的情感分析也得到了关注。[张博等, 2015] 将典型相关性分析引入情感分析迁移学习，基于特征映射迁移学习的思路，在保持各领域特有特征与领域共享特征相关性的基础上选择合适的基向量组合训练分类器，使降维后的相关特征在领域间具有相似的判别性，有效提高跨领域情感分类准确率。[Gui et al. 2015a] 应用类噪音估计算法，对跨语言迁移学习中负面样本进行检测和过滤，

提高了在目标语言情感分析的性能。[25]

随着基于深度学习的分布式表示算法的提出，能够利用神经网络获得对应的词语复合、句子复合、段落复合的分布式表示，在分布式表示学习的基础上进行文本情感分析的方法正在成为最近的研究热点。[Socher et al.2013] 利用递归神经张量网络模型对 Stanford 情感树库中标注了句子分析树和节点情感标签的句子进行复合表示学习。并应用该模型对输入句子进行逐个复合节点的情感判断，最终在根节点上得到整个句子的情感。[26] [Kim et al.2014] 将卷积神经网络进行部分改进后用于文本分析任务。实验结果证实了卷积神经网络在文本分析任务上的有效性。在这个的工作之后，[27][Cao et al. 2015] 利用卷积神经网络提取句子特征的向量表示，使用 SVM 作为分类器取代卷积神经网络 (CNN) 中的全连接层进行情感分类，提高了分类性能。[梁军等, 2015] 将长短时记忆模型 (Long Short Term Memory, LSTM) 模型扩展到基于树结构的递归神经网络 (RNN) 上，用于捕获文本更深层次的语义语法信息，根据句子前后词语间的关联性引入情感极性转移模型，到达了较好的情感分类性能。在文档级别的情感分类任务中，[28][Tang et al. 2015a] 提出了一种自底向上学习文档级向量表示的模型。该模型首先使用卷积神经网络 (CNN) 或长短时记忆网络 (LSTM) 学习句子级向量表示，随后，利用 gated RNN 将句子语义和联系自适应编码进文档的向量表示中。利用不同级别的文本的复合表示学习，提高情感分类的性能。针对文本情感分类任务通常面对的样例数据不平衡的问题，[15][Xu et al. 2015] 提出了一种利用词向量表示构造平衡的训练数据集的方法。在依次构造词向量和句子向量表示后，使用 SMOTE 算法生成少数类别的样例，并最终构造出均衡的训练数据，有效提高了文本情感分类的性能。

1.3.2 社交媒体文本情感分析研究现状

文本情感分析在社交媒体应用主要集中在在线新闻评论的观点挖掘和微博 (Twiiter) 文本情感分析，有关在线新闻评论的研究工作主要集中在信息检索方面，例如过滤，排序和评论摘要 (Potthast et al., 2012)。对比于上述信息检索相关技术，在新评论的观点挖掘研究方向上探索相对较少。随着微博的风靡, 与之相关的研究得到学术界和工商界的广泛关注。中文微博情感分析作为微博分析的重要基础任务，吸引了很多研究者的关注。以下分别介绍文本情感分析在两者的应用与研究。

对新闻评论的情感分析主要集中在极性检测和情感检测上。现有的研究大部分使用有监督的学习方法。在 (Zhou et al,.2010) 对比了不同的特征在情感分析上的作用，Chardon et al. (2013) 探讨了使用话语结构预测新闻反应的作用。在（Zhang

et al.，2012）中，提出了一种用于标记情绪（如悲伤，惊喜和愤怒）的元分类器。在他们的方法中，作者在评论中使用了两个异构信息源：基于内容的信息和情感标签。Jakic（2011）提出了一种自动预测新闻反应中情绪极性的方法。在这项工作中，作者使用了领域基础知识和迁移学习，得到了 Twitter 数据迁移训练的分类器。moreo 等人（2012）提出了一种基于词法的方法，可以适应不同的领域。在他们的工作中，他们使用 WordNet 关系设计构建了一个结构化的词典。Zhao et al. (2010a) 提出一种利用评论的聚类对评论分类的无监督的方法，检索每一个评论的关键句子并提取其中的命名实体作为评论的目标。

从读者调查 (Pang and Lee, 2008; Liu and Zhang, 2012; Mohammad) 和最近共享的比赛任务 (Wilson et al., 2013a; Rosenthal et al., 2015) 看，twitter 的情感分析研究很多。witter 情感分析比起正式本文有更大的挑战。它通常短且不正式，包含很多特殊的标记标签和表情符号和俚语，字母大小写也不一致。另一个问题是它倾向于不遵守语法规则，包含很多错误的拼写和很多单词的缩写。因此之前一些研究提出的了针对 Twitter 本文情感分析的方法。(Kiritchenko et al., 2014b) 提出根据 Twitter 文本短且非正事的特点，调整一系列表层特征提取方式，比如存在/没有积极和消极的表情符号、标签、大写的字幕和重复了字母的单词（比如 sweeettt）。最近几年，在分析对选举人 Twitter 政治情感、情绪和目的上研究热情很高 (Mohammad et al., 2015)，(Golbeck and Hansen, 2011; Conover et al., 2011a) 研究了如何确定政治联盟，(Maynard and Funk, 2011) 研究了如何识别有争议的问题和政治观点，(Conover et al., 2011b) 研究了如何检测选民两极分化的数量，（Tumasjan et al., 2010; Bermingham and Smeaton, 2011; Lampos et al., 2013) 研究了预测选举的投票意向或结果。更新的一个工作是 Lampos et al. (2013) 分析从英国和奥地利并成功在 300 多个跨两国民意调查中预测投票意向。

1.3.3 文本立场分析研究现状

上文论述了文本立场分析与文本情感分析有着本质的区别，文本立场分析更加关注文本反应出作者对于某一特定目标主题所持的立场和倾向。立场分析需要结合目标主题和情感信息，这比单独考虑文本的消息更加具有有挑战，对模型的建模能力也有更高的要求，现有立场分析的研究主要集中在国会辩论和网上辩论，这些领域的辩论作者会提出或者表现出清晰的立场倾向，语法和论述结构也相对固定化。但是对于一些用户主导的且表达方式更加随意内容，例如微博、Twiiter、商品评论的立场分析的研究也相对较少。

现有立场分析的研究的主要基于有监督的学习，在（Somasundaran 和 Wiebe，2010）的研究中，建立了论点触发词典，词典用来定位与抽取不同的论点，这些提取的论点、情绪表达、以及其目标作为立场分析分类器的特征。作者的实验表明单独用词频做特征比外加其他句法结构依赖性能效果并不会差很多，说明了在作者的任务上，其他特征对立场分析的影响相对较少。Hasan [20] 在 Anand 使用的特征集的基础上, 使用条件随机场 (Conditional Random Fields, CRFs)

标注用户交互序列特征、整数线性规划 (integer linear programming,ILP) 建模作者意识形态约束特征, 在 SVMs 分类器上的表现取得了最高 10% 的提升。通过两个支持立场语言的集合，Faulkner (2014) 研究了文档级别在学生的文章中立场分析。Hasan（2013）等提出了连续的评论之间不是完全独立的假设，连续的评论可以把问题定义成一个序列标注问题。Ahmed（2010）等提出一个该版本的主题模型算法（Latent Dirichlet Allocation LDA），作者把每一个词看成是立场倾向和主题的相互结合。Tutek 使用随机森林 (Random Forest,RF)、迭代决策树 (Gradient Boosting Decision Tree , GBDT)、逻辑斯蒂回归 (Logistic Regression,LR) 和支持向量机（Support Vector Machine SVM）四种基础的分类模型, 构造多组语言学特征, 作者利用了集成学习的思路，以 F-measure 作为集成学习的优化目标，线性组合四中基础模型的分类概率。其实验证明了基于多种基础分类的输出线性组合能明显的提高立场分类的效果。Sobhani 的研究表明, 一段文本表达的立场和文本的情感分析存在一定的关联。若将文本的情感作为文本立场分析的特征时，能显著的提升文本立场分析的性能。Rajadesingan and Liu (2014) 研究用户级别的立场分析，作者提出了如果多个 Twitter 用户转发同一对有争议的话题，那这些用户很大可能拥有同意的立场。

在自然语言处理领域中，传统的基于特征工程的方法能取得较好的效果，然而对文本提取特征词的特征工程需要大量的人力劳动和先验知识。而且基于传统的词向量的表达方式有语言缺乏关联，高纬特征稀疏特征表示，维度灾难等缺陷。Mikolov（2013）提出了 Word2Vec 模型，解决了词向量训练速度慢，效率低的缺点。其利用了 CBOW 和 Skip-Gram 的两种语言模型，且创新性的提出了 Hierarchical Softmax 和负采样的词向量加速方法。为后续深度学习模型能在自然语言处理任务上打下了夯实的基础。有关文本立场分析的研究也开始关注能自动提取特征的深度学习。Wan（2016）等利用多卷积核文本 CNN 的模型对 Tweet 文本进行有监督的立场分析，作者利用词窗口大小 3,4,5 的卷积提取文本中的特征，这算方法借鉴了自然语言处理中的 N 元组词（N-gram）的思想。此模型在 Semeval2016-TaskA 取得较好的成绩。Zarrella [32] 使用迁移学习的方法, 首先从大量的不标注 Twitter 数

据中，选取了词频高于 100 的词汇，然后用 Word2vec 模型预训练好了每一个词的 256 维度的词向量，后面通过词向量相似度选取比较关键的以 # 为前缀主题标签，通过训练神经网络预测主题标签。后通过迁移学习的思想对网络结构进行微调来达到立场分析的目的。实验结果表明, 外部无标注数据的使用可以给有监督学习提供一些帮助，提高有监督学习的性能。

上述主要叙述了基于有监督的立场分析方法，虽然有监督的方式可以准确拟合训练集中，构造出对训练集有显著效果特征, 但是标注大量有类标的训练集需要大量的人工成本，模型泛化能力也相对较弱，技术的应用场景也十分有限。为解决上述的缺点，研究人员开始展开对文本立场立场分析的无监督学习和弱监督学习的研究。Johnson [24] 等通过基于不同方面的特征，构建 6 个局部弱监督基分类器。基于这些弱监督分类器在概率软逻辑 (Probabilistic Soft Logic,PSL) 算法的结合下，组成一个全局的弱监督分类模型。此模型在有关美国的 32 名政治人物的 Twitter 文本的立场分析任务中取得较好的性能。Augenstein [36] 使用基于词袋的自动编码机（Auto-Encoder）学习文本的特征表示, 并将学习得到的特征用于有监督的分类器中, 解决了训练数据缺失的问题。

1.4 本文的主要研究内容和组织结构

本文主要研究基于深度学习模型并对社交媒体中的文本进行立场分析的方法。

第 2 章 文本立场分析相关技术概述

2.1 引言

本章概要介绍立场分析及其相关技术。首先从目前研究相对成熟的文本情感分析入手，分布从传统的基于规则、机器学习、深度学习分别讨论文本情感分析技术。由于本文主要以深度学习的方法解决社交媒体中立场分析，所以单独详细分析深度学习在文本情感分析的研究。随后作为本文的重点研究对象详细介绍了分别基于机器学习和深度学习模型的文本立场分析技术相关研究本章总结了各项研究工作的特点，在分析优缺点的基础上引出本文的后续研究

本章 2.2 节介绍情感分析的相关研究，2.3 节介绍立场分析相关研究，2.4 节着重介绍基于深度学习模型的立场分析研究。

2.2 文本情感分析相关技术概述

文本情感分析，指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息。通常来说，情感分析的目的是为了找出说话者，作者在某些话题上或者针对一个文本两极的观点的态度。这个态度或许是他或她的个人判断或是评估，也许是他当时的情感状态，或是作者有意向的情感交流。文本的情感分析是自然语音处理的重要研究内容之一，且其具有重要的科研价值与商业实用价值，吸引了大量的研究人员的关注。研究人员从不同的角度和不同方法对文本情感分析展开了研究。本节将从基于情感词典、机器学习和深度学习的三个方向分别概述近年来情感分析的研究进展。

2.2.1 基于情感词典的文本情感分析相关技术

基于情感词典的文本情感分析是早期研究人员的成果，其相应的模型建立在情感词典和语言学的规则基础上。由于模型的解释性好，需要计算资源较少，成为早期研究文本情感分析的主流。情感词典作为文本情感分析的重要组成部分能给文本情感分析提供重要的特征信息。情感词典的构造通常有语言学领域的专家完成。例如现有先对成熟的情感词典有 WordNet、HowNet、大连理工大学中文情感词汇本来库等。基于情感词典的文本情感分析的技术步骤通常先匹配原文本中与

情感词典相对应的情感表达特征词，然后根据各特征词的表达方式综合计算其每一个特征词的情感得分，最好结合整个文本的情感得分总结文本的情感倾向。

Taboada【】在原来情感词典的基础上进一步考虑了词语的词性，结合情感词和词性综合给出情感的倾向得分。该情感分析模型包含一个语义指向计算器（Semantic orientation calculator，so－cal），这个计算器首先抽取出文本中的形容词、动词、名词以及副词等情感方位词，然后结合各种情感方位词计算原来文本的情感指向，模型结合的情感指向和强调、弱化、否定等转移的价位得到文本最终的情感倾向。作者通过一系列的实验证明了基于情感词典和此种转移规则的模型具有很强的鲁棒性，在跨领域的的文本情感分析上也有良好的表现。孙建旺【】等提出结合情感词典和机器学习两者的优势来解决微博情感分析的问题，利用微博多层次结构对微博文本进行特征降维。此外，由于微博包含多种颜文字，表情符等特点，设计了对颜文字和表情符的情感计算方法，其实验证明了加入表情符等特征，对微博的情感分析效果得到了提高。

基于不同的上下文可能决定某些情感词的特点，具有一定的领域相关性。例如“高”在“质量高”的上下文中表达的是正面的情感倾向，但是如果在“消费高”的上下文则表达负面的情感倾向。Bollegala【】等人结合了不同领域对情感词的表达特点构造了领域相关的情感词典。实验证明结合领域知识的情感词典能在相对于的领域取得更好的效果。Li【】提出一种相关领域自适应情感词的框架，能同步从标注训练语料中提取出的主题词和情感词，并进一步通过分析标注语料中主题词和情感词的关系来推导出未标注语料中与主题相关的情感词。

总体来看，基于情感词典/规则和知识库的情感分类准确率较高，但由于情感词典和常识库规模的限制，覆盖率较低。同时此类方法对分词、词性标注、规则匹配等的准确性要求较高，系统内部错误传递影响较大。[30]

2.2.2 基于机器学习的文本情感分析相关技术

随着机器学习成功应用于其他领域的快速发展，对于文本情感分析的问题，大量的研究人员开始开展基于机器学习的文本情感分析的研究。基于机器学习的文本情感分析方法，首先通过特征工程抽取文本情感分析特征，然后通过抽取出来的特征词用机器学习能理解的数值表达文本。通过人工标注建立起特征表示数据和情感类标对的训练数据，通过各种已有的机器学习模型（支持向量机、朴素贝叶斯、逻辑回归、最大熵模型等）提取出训练集中特征和类标之间的映射关系的模型。Sida Wang【42】等利用 N 元词组（N-gram）对文本情感进行建模，模型结合

了朴素贝叶斯与支持向量机的两个模型。首先利用朴素贝叶斯的思想，计算每一个词组的对数计数概率 r，公式如下：

$$r = \log(\frac{\frac{p}{\|p_1\|}}{\frac{q}{\|q_1\|}})$$

$$p = \alpha + \sum_{i:y^{(i)}=1} f^{(i)}$$

$$p = \alpha + \sum_{i:y^{(i)}=-1} f^{(i)}$$

其中 $y^{(i)}$ 为训练实例 i 的类标，$y^{(i)} \in \{-1, 1\}$。其中 $f^{(i)}$ 为训练实例 i 的特征向量，$f_{(i)} \in R^{\|V\|}, V$ 为特征集合 α 为平滑因子。

于上述 r 的计算公式可知，从训练语料中可以计算出每一个词语对于不同情感的倾向大小，所以利用我们已经计算的每一个词的 r 值，可以得到文本的特征表示，特征表示后的文本可以作为支持向量机的输入，通过支持向量机可以抽取出训练集中有关文本情感分析的模式。

Pang[8] 等研究者创新性的把文本主题分析的技术迁移应用到文本情感分析中, 文本的话题分类主要根据与话题相关的主题词决定，但是表达情感的方式更加的多样话，需要考虑的因素更多。Pang 把文本的情感分析看成一类特殊的主题分析，使用了在有监督学习上泛化能力较好的支持向量机、朴素贝叶斯、最大熵模型三种基础的分类模型。选用的分类特征为一元词组（Unigram）、二元词组（Bigram）、词性分析（POS）、形容词位置信息等。此研究通过组合特征和模型的交叉验证表明，三个分类器组合任意一个特征特征的性能都比基线模型要好，在有关电影影评的数据集上，一元词组（Unigram）结合支持向量机的模型取得了良好的效果。但是此研究实验也论证了文本的情感分析的性能还是和文本主题分析存在较大的差距，同时也佐证了文本情感分析对模型和特征也有更高的要求。

为了减少文本中无关的客观信息对文本情绪分析的干扰作用，Pang 和 lee 对上述基于机器学习的文本情感分析模型进行了有正对性的改进，规避了文本客观消息的干扰，使模型更加专注于文本的主观信息。作者他们创新性把原来的文本情感分析问题转换成以各字句链接图中最小割问题，应用了挖掘图中的最小割的分类器来寻找对情感分析有用的主观表达的句子，从而屏蔽掉客观消息的干扰。此研究的实验也论证了剔除客观信息的文本情感分析模型的性能得到显著的加强。

2.2.3 基于深度学习的文本情感分析相关技术

深度学习由于其更复杂的模型和更多的参数，比起以往的的机器学习方法在海量数据集上更有优势。且深度学习具有能够以端到端的形式构建模型的优势，不再需要人工筛选和总结大量特征，所以得到了情感分析研究者的关注。Socher 等研究者为满足深度模型构建的需要，首先组织标注了斯坦福情感树库（Stanford Sentiment Treebank，SSTB）。斯坦福情感树库由 11,855 个电影评论句子组成，共包含 215,154 个不同短语，其中任意短语构成的节点和其他叶子节点均被标注为五类情感（强正面、正面、中性、负面、强负面）中的一个。Socher 提出使用语法树和词汇向量表示任意长度的短语输入的递归神经张量网络，且其使用相同的张量组合函数来计算根节点向量，可得到句子中任意短语的情感向量表示。该模型可以利用树形结构捕捉情感变化和否定侧的作用范围，对于转折结构中情感表达识别同样具有很好的效果。实验证明，RNTN 模型在五分类的情感分析及二分类（“正”、“负”）的情感分析中都取得了历史最好成绩。

最简易的文字编码方法就是 one-hot representation，向量长度为整个语料库中词的总个数。向量的分量只有一个 1，其他全为 0，1 的位置对应该词在词典中的索引。但这种词向量表示有一些缺点，如容易受位数灾难的困扰，且不能很好地刻画词与词之间的相似性。另一种词向量是 Distributed Representation，它最早是 Hinton 于 1986 年提出的，可以克服 one-hot representation 的上述缺点。其基本想法是：通过训练将某种语言中的每一个词映射成一个固定长度的短向量（这里的“短”是相对于 one-hot representation 的“长”而言的），所有这些向量构成一个词向量空间，而每一向量则可视为该空间中的一个点，在这个空间上引入“距离”，就可以根据词之间的距离来判断它们之间的（语法、语义）相似性。bengio[33] 等人提出用神经网络的方式建立二元的语言模型，把词映射为低纬度稠密的词向量，并用词向量之间的距离来衡量词语之间语义的相似性。Mnih[34] 等人提出基础层次 Log-Bilinear 模型来训练神经网络中的语言模型。Mikolov[??]2013 提出了 Word2Vec 模型，解决了词向量训练速度慢，效率低的缺点。其利用了 CBOW（Continuous Bag-of-words Model）和 Skip-Gram 的两种语言模型。其中 CBOW 的思想是利用词语的上下文词的信息来预测该单词。而 Skip-Gram 则采取一种和 CBOW 相反的策略，用中间的词的消息来预测上下文的词。

Mikolov 创新性的提出了 Hierarchical Softmax 和负采样的词向量加速方法, 为后续深度学习模型能在自然语言处理任务上打下了夯实的基础。对已经训练好的词向量，通过 PCA 等降维方法可在低纬度空间内实现可视化。

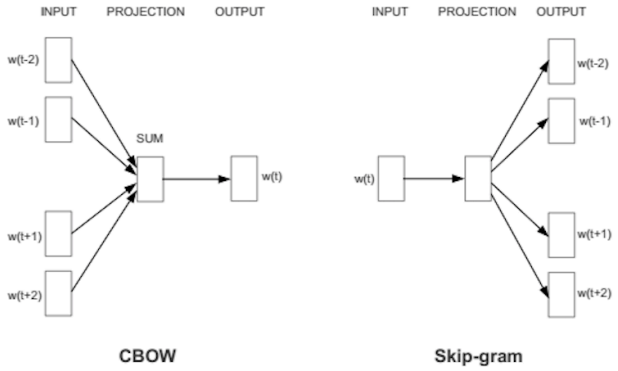

图 2-1 Skip-gram 和 CBOW 模型

循环神经网络 (Recurrent Neural Networks-RNN) 已经在众多自然语言处理任务上取得了巨大成功。不同于传统过得前向反馈神经网络的同层节点无连接层于层之间节点有连接，循环神经网络引入了定向循环，可以处理序列数据。RNN 中最大的缺陷是后面时间的节点对于前面节点的感知力下降，当网络深时训练效果下降。LSTM 可以解决这一问题，目前 LSTM 是 RNN 中使用最广泛最成功的模型。

卷积神经网络不仅在图像处理上表现优异，在文本分类上同样表现不俗。kim[]提出的多卷积核文本分类 CNN 模型。模型第一层为预先训练好的词向量或者是随机初始化参数的词向量 Embedding 层，然后接一个宽度和词向量维数相同，长度通常为 3、4、5 的多个卷积核的卷积层，卷积层后是是较少参数的池化层，最大值池化层在分类的效果较好。池化层后的输出为文本提取出来的特征向量，最后连接一个全连接作为分类层。实验证明 CNN 在情感分析上能较好的性能。

2.3 文本立场分析相关技术

文本立场分析与文本情感分析有着本质的区别，文本立场分析更加关注文本反应出作者对于某一特定目标主题所持的立场和倾向。立场分析需要结合目标主题和情感信息，这比单独考虑文本的消息更加具有有挑战，对模型的建模能力也有更高的要求。

2.4 文本立场分析相关技术

作为一项特殊的情感分析任务，立场分析问题主要是在给定了目标的前提下，判断这个文本的立场是“支持”或“反对”。作者在文本中评价的目标不一定是立场分析给定的目标，也可能立场分析给定的目标并没有直接出现在文本中。即使

作者在文本中对目标/实体的态度是积极的，但推断出来的结果可能是作者对给定目标持反对的立场。

具有强大学习能力的基于特征的机器学习算法能够充分学习到文本的语法、语义等特征，因此受到了研究者的广泛关注。本节将介绍有监督的立场分析方法和若监督的立场分析方法。考虑到深度学习模型的立场分析研究逐渐增多且有成为主流的趋势，将在 2.4 小节单独介绍基于深度学习模型的立场分析研究状况。

2.4.1 基于有监督机器学习的文本立场分析技术

Anand 等作为立场分析问题较早的研究者之一，抽取了网络论坛上 4837 篇分属 14 个不同话题的讨论，并针对此语料，构造了 10 种不同的特征。这 10 种特征包括一元/二元单词、文本长度、重复标点、线索词、句法依赖、语义依赖、广义依赖、上下文特征和 LIWC（Liunguistics Inquiry Word Count）等。分别利用 RIPPER 和朴素贝叶斯两种算法验证了立场分析在不同特征集合下的效果。实验证明，应用了上下文特征可以有效提升实验效果，在不同子话题中统一分类模型表现有较大差异。对于 Anand 的研究，Hasan 等研究者认为他们忽略了文本间意识形态和用户交互这两种对实验性能具有很大影响的约束。Hasan 在研究中的文本序列模型是以相邻文本减的用户交互约束建立的，并利用了 CRFs 来解决序列标注问题。该研究为每位文本作者的意识形态以 ILP 建立基于阔话题领域、作者的意识形态约束模型。并且在实验中得到了比 Anand 正确率上 2.9% 到 10% 不等的提升。

社交文本以推特为例，具有长度较短且语言非正式等特点，Zhang 针对这些特点提出了两部的学习系统。将社交媒体文本的三个立场“支持”、“反对”和“其他”转化为两次两分类的任务。第一步进行文本相关性检测，可以将无关文本作为“其他”类型检测出来；第二步使用立场倾向检测模型，对文本进行“支持”“反对”二分类。该研究在传统语言特征外还应用了 LDA（Latent Dirichlet Allocation）生成的话题相关性、主题只是、情感词汇、词嵌入和表情符等多种不同的特征组合，为每一组子话题数据分别使用线性回归模型（Linear Regression，LR）简历立场分析模型。并且考虑到不同话题数据分布具有差异，对这来年各个子模型验证了不同超参数和特征集的组合方式。最后用实验证明了对不同子任务和子话题使用不同的特征组合的方式是有效的。

除了不同过得特征组合可以起到提升立场分类效果的作用，构建有差异的分类器组合也同样可以提升模型的立场分类效果和泛化能力。Xu 等【】提出了选用该段落向量、浅层语义分析、LDA 和其他语言学特征等，在线性 SVMs、RBF 核

SVMs、Adaboost 随机森林和随机森林等分类其中测试分类性能。并且该模型采用了对及分类器线性组合的方式提升模型表现，具体如公式 2-1 所示。

$$p(C|x) = \sum_{i=1}^{m} w_i \cdot p_i(C|x) \tag{2-1}$$

式中 $p_i(C|x)$ 为文本 x 在第 i 个分类器中被预测成立场类别 C 的概率；

w_i 为线性模型中及分类器第 i 个基分类器的权重

$p(C|x)$ 为文本 x 在立场类别 C 中的概率

实验证明，该模型可以提高立场分类效果，并且可以直接被扩展到相似任务中。

2.4.2 基于弱监督机器学习的文本立场分析技术

弱监督学习是一种基于噪声训练数据的半监督学习方式。不同于以往基于可靠训练数据的有监督学习方式，弱监督学习基于的是不严密的假设生成训练数据。在训练数据中更可能含有错误和大量噪声。但该方法可以避免人工标注的高昂成本，且能在新语料、新话题中具有较好适用性。

Johnson 等【】研究者提出了基于若干弱监督局部分类器的联合模型。每个局部分类器使用的训练语料为少量有立场和框架维度标注的种子集合，随后使用概率软逻辑框架组合各弱监督局部分类器。弱逻辑框架使用的训练数据是弱监督局部分类器的信息，之后利用合页-马尔科夫随机域 (hinge-loss Markov random fields，HL- MRFs) 的图模型关系表示建立用于立场预测的规则。实验证明该模型能较准确的预测推特政治任务立场。

Ebrahimi【】提出了使用关系自扩展方来实现弱监督学习的立场分类法扩展种子训练集。定义了三种立场相关约束：相似文本表达的立场相似、具有朋友关系的作者立场相似和文本作者对话题的立场相似。Ebrabimi 为了匹配含噪音的少量标注集合，首先使用若干短语模式，然后利用基于统计关系学习的合页马尔科夫随机域模型标注文本立场，最后使用结合词典、多元词组和情感类型特征的线性和 SVMs 训练有监督的立场检测模型，并在 SemEval 数据集上表现良好。

Dias【】提出了使用启发式规则的方法自动标注 Tweet 文本，可以解决弱监督学习中预料标注的问题。该方法可以达到两重目的，既可以使用监督学习算法自动创建训练语料库开发预测模型，又可以补充预测立场检测模型。Dias 构建了 7 条启发式规则构建训练语料，分析了六个不同的立场分析任务，取得了可观的成果，加权 F 值从 52% 提升到了 67%。

2.5 基于深度学习的立场分析技术

近年来深度学习不仅在计算机视觉、语音识别等领域取得了卓越的成果，在自然语言处理中也有越来越多的研究者开始采用深度学习模型。不同于传统机器学习方法的手动特征工程，深度学习模型采用端到端的分类模式，能够自动抽取在参数学习中最优的特征表示。本节介绍深度学习模型在自然语言处理中的进展和立场分析方法上的应用。

Mikolov【】提出了一种基于循环神经网络的语言模型（Recurrent Neural Network Language Model, RNN-LM）。该网络具有三层结构，除了输入输出层还有上下文表示层，上下文表示层的信息拼接了上一时间和当前时间的两个信息，因此可以将全部历史信息保存在低纬度的向量空间中，大大增强了语言建模能力。实验表明，与已有的退避模型（Backoff Language Model）相比，使用多个 RNN-LM 可以减少约 50% 的困惑度（perplexity）。在“华尔街日报”任务上降低了 18% 的错误率。Sundermeyer 在【】为解决 RNN-LM 模型在反向误差传播函数时梯度消失/爆炸带来的巨大影响，采用了结合长短时记忆单元的循环神经网络（LSTM），该单元包含三个门：输入门、输出门和遗忘门，这样的结构可以解决梯度缩放问题且计算代价小。该模型在标准语料库上的困惑度指标比 RNN-LM 模型低 8%。

Zarrella 等【】研究者同样使用了 LSTM-RNN。并且用 word2vec skip-gram 方法训练了单词和短语的词向量。采集了 218,179,858 条 Twitter 文本，训练过程迭代 100 次，为文中出现的 537,366 个词汇和短语训得到 256 维的词向量。然后通过主题标签预测辅助任务来学习句子向量，然后被微调用于标记样本的立场检测。实验结果显示，该方法在 5 个话题文本立场检测任务中 F 值为 0.678。

Yu【】提出了基于双向 LSTM-RNN 的模型，该模型包括词嵌入输入层、卷积层、长短时记忆单元层、双曲正切输入层和全连接分类层。卷积层的作用是提取多元词组的特征，长短时记忆单元曾的作用是学习潜在全局语义，池化层的额作用是减少参数并归纳句子表示。组合上述的知识表示用最后的双曲正切函数。实验证明此模型在中文立场分析任务中表现不错。

现有的立场分析任务中，普遍忽略了预制话题的作用，只把立场分析看做是句子级别的简单分类任务。Augenstein【】为了解决这一问题，在立场分析弱监督任务中使用了双向条件编码的长短时记忆模型，模型如图 **??** 所示。模型中使用了固定长度编码的 LSTM 单元，使用该向量初始化第二个 LSTM 单元，Twitter。在 SemEval 2016 任务 6 Twitter Stance Detection 语料库上获得了更高的性能。

Wei【】为了解决在弱监督立场分析中标注数据缺失的问题，使用了两步框架。

构建粗糙的二分类器，使用他们定义的 softmax 层在二分类训练数据中执行三分类任务。第一步利用有清晰倾向性的词汇和表情标注支持和反对两种立场，第二部将“支持”立场和“反对”立场概率差绝对值小于阈值的文本标注为第三类“中立”。以弱监督标注文本为基础，使用与有监督学习相似的深度学习模型得到的预测结果的方法被实验证明有效改善了训练数据在弱监督立场分析中缺失的问题。

近年来很多研究者开始关注在计算机视觉研究中较火的“注意力机制”并将其用于自然语言处理任务中。Bahdanau 等【】第一次在机器翻译领域尝试“注意力机制”。该研究的编码器是双向 RNN 模型，解码器是 gate-RNN 模型。注意力机制被使用在解码器中用以发现编码器对当前翻译输出最有用的隐藏状态。注意力权重矩阵决定当前的翻译内容。除机器翻译之外，在语言模型和智能问答【】、图片描述文字生成【】、和句法树生成【】等领域也得到了广泛的应用。

2.6 本章小节

本章首先介绍了情感分析的常用技术和研究现状，然后从基于手工特征工程的机器学习方法和端到端的深度学习模型来年各个角度详细描述了立场分析检测的现有研究。基于特征筛选和分类器集成学习仍然是机器学习领域的主流，基于 RNN/LSTM 和 CNNs 等模型的工作是在深度学习领域上的主要方案。如何有效利用外部信息学习出更好的文本表示在现有的研究中仍然是一个重要的课题；其次，特有的预制话题包在分析任务任务中对最后的预测也起到了十分关键的作用，立场分析就有工作的重要方向就是如何将预制话题短语信息更好地利用上。

第 3 章 基于条件编码长短期记忆社交媒体文本立场分析

3.1 引言

当前社交媒体文本立场分析的研究中，研究人员的研究方向主要集中在如何提取社交媒体文本有效的分类特征。忽略了文本立场分析一种重要的出发点是文本基于某个特定的目标，若原文本脱离了特定目标，文本立场分析与情感分析将无差别。基于这个出发点，本章提出了基于条件编码长短期记忆神经网络的模型，通过以条件编码的形式引入文本的目标消息，使立场分析的效果得到显著的提升。本章研究基于条件编码长短期记忆的社交媒体文本立场分析方法，通过以不同形式给模型接入文本立场的目标消息，表明了接入目标信息对文本立场分析有明显提升效果。通过在 SemEval2016 英文立场分析数据集和 NLPCC2016 中文立场分析数据集的实验，论证上述结论。

本章的各节结构如下:3.2 节介绍条件编码长短期记忆神经网络模型;3.3 节介绍基条件编码长短期记忆在文本立场分析;3.4 节为本章实验和结果分析; 最后一节为本章小结。

3.2 条件编码长短期记忆神经网络模型

循环神经网络 (Recurrent Neural Networks-RNN) 已经在众多自然语言处理任务上取得了巨大成功。不同于传统过得前向反馈神经网络的同层节点无连接层于层之间节点有连接，循环神经网络引入了定向循环，可以处理序列数据。RNN 中最大的缺陷是后面时间的节点对于前面节点的感知力下降，当网络深时训练效果下降。LSTM 可以解决这一问题，目前 LSTM 是 RNN 中使用最广泛最成功的模型。Rocktaschel[???] 在 2016 年在句子之间的文本蕴含识别（Recognizing textual entailment RTE）的研究中提出了条件编码的思想，其论证了在文本含义识别任务上，条件编码比单独编码更能抽取两个句子之间的信息。结合文本立场分析的也有文本和目标需要同时考虑特点，把借鉴条件编码的思想来解决文本立场分析的任务。

3.2.1 基于 GloVe 的词嵌入模型

词的表示是自然语音处理中一个基础且十分重要的任务，大量的研究人员投入到词表示的研究中。早期的有词的表示方式为 one hot encoding。每个词独占一个维度，每个词向量有一个维度是 1，其他维度为 0，词向量的维度是所以单词的的长度。One hot 编码的特点是假设所有的单词互相独立，这是一个很强的假设，显然在有些任务中并不合适，如词语相似度方面，dog 和 cat 的相似度应当比 dog 和 not 高，但是在 one hot 编码中他们相似性一样。one-hot 编码词表示有一些缺点，如容易受位数灾难的困扰，且不能很好地刻画词与词之间的相似性。词嵌入表示模型能很好的改善 one-hot 编码的缺点，词的嵌入模型用稠密且固定长度的向量表示每一个词，而且相似的词具有相似的词向量表示。基于这点 Mikolov[??]2013 提出了 Word2Vec 模型，解决了词向量训练速度慢，效率低的缺点。其利用了 CBOW（Continuous Bag-of-words Model）和 Skip-Gram 的两种语言模型。其中 CBOW 的思想是利用词语的上下文词的信息来预测该单词。而 Skip-Gram 则采取一种和 CBOW 相反的策略，用中间的词的消息来预测上下文的词。GloVe(Global Vectors for Word Representation) 是斯坦福大学发表的一种词嵌入模型,GloVe 尝试借鉴 NNLM 和 word2vec 的优势来弥补旧方法的劣势，取得了不错的效果。该文发表于 word2vec 之后，其方法内核比较朴实和简单，官方实验中，GloVe 是略胜 word2v 一筹。

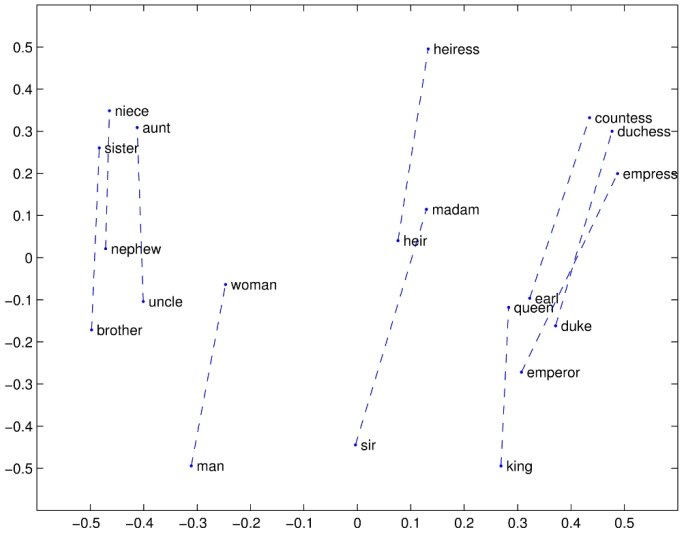

图 3-1 GloVe 可视化词向量

GloVe 结合了基于矩阵分解的词嵌入表示和基于语言模型（例如 Word2Vec）的词嵌入表示。基于矩阵分解具有训练快，容易实习的优点，但是生成词向量语义消息十分有限。基于语言模型的词嵌入表示则具有更多语言学层次的支持，在更

多自然语言处理任务上表现更好,但是模型关注的上下文特征较少，忽略了全局的信息。GloVe 发明的初衷，就是想结合两者的长处，建立一个充分利用统计量的更好训练的适用程度更广的词嵌入模型。具体模型建立公式如下

$$F(w_i, w_j, w_k^c) = \frac{P_{ij}}{P_{jk}}$$

其中，取 $word_i$ 的出现次数为 X_i，定义 $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ 表示在 X_i 的上下文下 $word_i$ 的出现几率, F 则是某一种能够实现我们需求的变换。w_i, w_j 是实数空间下的 $word_i, , word_j$ 的词向量，w_k^c 也是实数空间下的 $word_k$ 的上下文词向量，其作用类似 word2vec 中的上下文向量。为了精简计算引入词向量的线性加减和点乘计算

$$F((w_i - w_j)^T w_k^c) = \frac{F(w_i^T w_k^c)}{F(w_j^T w_k^c)}$$

GloVe 每个词涉及到两个词向量，一个词语本身的向量 w_i，一个词的 context 向量 w_i^c。最初这样设计，将词向量和上下文的向量分开，不用一套，是为了在更新参数的时候能够简单地套用 SGD。实验证明两个向量加起来最后起到的效果最好。后面英文的词向量用的是 GloVe 模型在大量的 Twitter 文本上训练的 100 维度的词向量，中文微博词向量是 200 维度的词向量。

3.2.2 长短期记忆神经网络模型

由于文本序列的通常具有较长的长度，导致神经网络的层数较多，而传统的递归神经网络解决序列问题经常会出现梯度消失的问题 (vanishing gradient problem) 与梯度爆炸问题（gradient exploding problem）。梯度消失问题和梯度爆炸问题一般随着网络层数的增加会变得越来越明显。出现的原因在于对神经网络参数进行链式求导的过程中，输出对于前面递归神经参数的倒数随着累乘激活函数的导数而接近于 0，以下图的反向传播为例（假设每一层只有一个神经元且对于每一层 $y_i = \sigma(z_i) = \sigma(w_i x_i + b_i)$，其中 σ 为 sigmoid 函数）

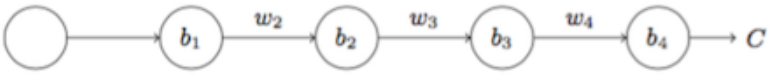

图 3-2 RNN 梯度消失

可以推导出

$$\frac{\partial C}{\partial b_1} = \frac{\partial C}{\partial y_4}\frac{\partial y_4}{\partial z_4}\frac{\partial z_4}{\partial x_4}\frac{\partial x_4}{\partial z_3}\frac{\partial z_3}{\partial x_3}\frac{\partial x_3}{\partial z_2}\frac{\partial z_2}{\partial x_2}\frac{\partial x_2}{\partial z_1}\frac{\partial z_1}{\partial b_1} \tag{3-1}$$

$$= \frac{\partial C}{\partial y_4}\sigma'(z_4)w_4\sigma'(z_3)w_3\sigma'(z_2)w_2\sigma'(z_1) \tag{3-2}$$

一般的非线性激活函数的导数都小于 1（例如 sigmoid 的导数最大值为 $\frac{1}{4}$），因此对于上面的链式求导，层数越多，求导结果 $\frac{\partial C}{\partial b_1}$ 越小，因而导致梯度消失的情况出现。长短期记忆（Long Short-Term Memory, LSTM）是一种缓解上述问题的间递归神经网络的变种，Hochreiter 在 1997 年首次提出了 LSTM 结构，2000 年 Gers 等人改进 LSTM 模型。

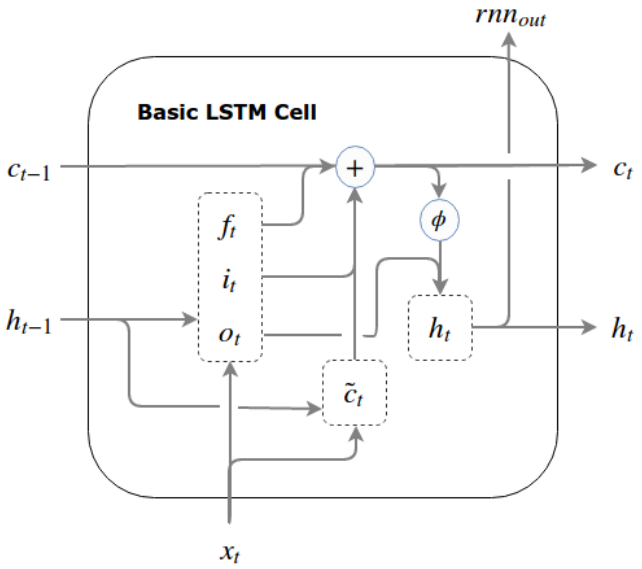

图 3-3 LSTM 单元结构

LSTM 模型提出了记忆存储格（memory cell）的结构，内部包含了遗忘门（forget gate）、输入门（input gate）和输出门（output gata）。各种门的作用在于调节记忆体在外部输入的情况下应该采取怎么的存储测量，具体门状态和记忆体内部参数的更新公式如下。

$$i_t = \sigma_g(W^i x_t + U_i h_{t-1} + b^i) \tag{3-3}$$

$$f_t = \sigma_g(W^f x_t + U_f h_{t-1} + b^f) \tag{3-4}$$

$$o_t = \sigma_g(W^o x_t + U_o h_{t-1} + b^o) \tag{3-5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{3-6}$$

$$h_t = o_t \odot \sigma_h(c_t) \tag{3-7}$$

其中 σ_g 为 sigmoid 的激活函数，σ_c, σ_h 为 thah 的激活函数，x_t 为输入向量，h_t 为输出向量，c_t 为记忆向量，W, U, b 是矩阵参数和向量参数。各个门的值是保持在 0-1 之间的向量。其中遗忘门向量 f_t 表示上一时刻的记忆体信息需要遗忘多少，输入门向量 i_t 表示有多少当前时刻输入信息需要加入到记忆体中，输出门向量 o_t 表示记忆体输出多少信息。

前已经证明，LSTM 是解决长序依赖问题的有效技术，并且这种技术的普适性非常高，导致带来的可能性变化非常多。各研究者根据 LSTM 纷纷提出了自己的变量版本，这就让 LSTM 可以处理千变万化的垂直问题

3.2.3 条件编码长短期记忆神经网络模型

Rocktaschel[] 等在句子之间的文本蕴含识别的研究中提出了条件编码长短期记忆神经网络模型，文本蕴含定义为一对文本之间的有向推理关系，其中蕴含前件记作 T(Text)，蕴含后件记作 H(Hypothesis)。如果人们依据自己的常识认为 H 的语义能够由 T 的语义推理得出的话，那么称 T 蕴含 H，记作 T → H, 作者提出的模型的结构是首先有一个 LSTM 模型编码 Text 消息，另一个不同参数的 LSTM 模型编码 Hypothesis。作者不是简单把两个特征向量拼接在一起，而是做了如下转换。把第一个编码 Text 信息的 LSTM 模型的记忆状态（Cell）保留下来，作为第二个编码 Hypothesis 的 LSTM 模型记忆状态（Cell）的初始值，此模型建立的了 Text 消息作为条件下的对 Hypothesis 的编码表示。

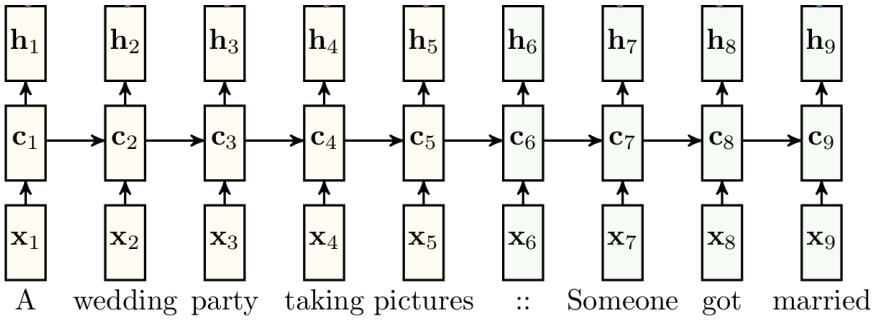

图 3-4 条件编码长短期记忆

如上图图表示所示“A wedding party taking pictures”作为我们的 Text 文本，“Someone got married”作为我们的 Hypothesis，其中 c_5 作为前一个 LSTM 的记忆体状态被当做编码 Hypothesis 的初始记忆状态。两个 LSTM 具体的状态转移公式如下：

$$[h_1\ c_1] = LSTM^{Text}(x_1, h_0, c_0) \tag{3-8}$$

...

$$[h_T\ c_T] = LSTM^{Text}(x_1, h_{T-1}, c_{T-1}) \tag{3-9}$$

$$[h_{T+1}\ c_{T+1}] = LSTM^{Hypothesis}(x_1, h_0, c_T) \tag{3-10}$$

...

$$[h_N\ c_N] = LSTM^{Hypothesis}(x_1, h_{N-1}, c_{N-1}) \tag{3-11}$$

$$c = tanh(Wh_N) \tag{3-12}$$

其中 $(x_1...x_T)$ 为 Text 的序列消息，$(x_{T+1}...x_N)$ 为 Hypothesis 的序列信息。h_0, c_0 为 LSTM 的初始化向量。

实验证明在文本蕴含任务上，条件 LSTM 模型比单独编码高 3.3%（从 77.6% 提升到 80.9%）的性能。这种条件编码能使 Text 的信息更好的流向对 Hypothesis 编码的 LSTM 模型，有了第一个 LSTM 模型传来的记忆状态，第二个 LSTM 模型能更好的编码 Hypothesis 的消息。

3.3 基于条件编码长短期记忆的文本立场分析

通过 Rocktaschel 在文本蕴含的任务的实验可知，在处理两个文本序列的编码任务上，条件编码长短期记忆神经网络比单独独立编码两个文本序列有更好的建模能力。在文本立场分析的任务上，有文本的信息和目标主题两个文本序列信息，我们可以借鉴条件编码长短期记忆神经网络在文本蕴含建模的方式，把文本信息和目标主题信息更好结合起来。在实验部分设计了多种文本序列信息的结合方式，通过实验证明了以目标主题文本作为条件编码文本信息的模型对文本立场分析有更好的效果。

本文后期实验将在 NLPCC2016 中文微博立场分析数据集和 SemEval2016 英文 Twitter 立场分析数据集，为较清晰阐述条件编码长短期记忆模型，以下简短的介绍下两个数据集的样例，具体的有关数据集的信息将会在下面实验部分做详细介绍。

例 1: 目标主题文本:" 深圳禁摩限电" 微博文本:" 支持深圳交警。电单车继续治理" 立场分析类标：“Favor”（持支持立场）

目标的文本主题有关“深圳禁摩限电”的主题的，而从微博文本“支持深圳交

警。电单车继续治理”中，我们可以知道微博的作者首先是赞同了深圳交警的行为，然后叙述了电单车需要得到继续的整治，从两个方法肯定了” 深圳禁摩限电” 这个主题目标的，因此给出的类标是“Favor”也就是持支持目标主题的立场。

例 1: 目标主题文本:”Hillary Clinton” Twitter 文本:”Hopefully Hillary Clinton gets cancer and dies before she gets the opportunity to embarrass our country any further “，立场分析类标“Against”（持反对立场）

译文:” 真希望希拉里克林顿得癌症然后死去，这样她就不再会有机会再让我们国家蒙羞了。”

目标的文本主题有关“希拉里克林顿”的主题的，这个 Twitter 文本是有关 2016 年美国大选，显然 Twitter 作者一直咒骂希拉里克林顿，希望她得癌症，不让她侮辱国家，可以看出作者有强烈反对主题目标“希拉里克林顿”，因此给出的类标是“Against”，也就是持反对目标主题的立场。

从上述的两个简单的样例可知，立本立场是有两个输入的，一个是立场主题例如“深圳禁摩限电”和“Hillary Clinton”。另外一个是立场下的文本“支持深圳交警。电单车继续治理”和”Hopefully Hillary Clinton gets cancer and dies before she gets the opportunity to embarrass our country any further “。在这通过中文微博阐述条件编码长短期记忆模型的建立。首先经过一些数据预处理和分词把主题目标”深圳禁摩限电“和”支持深圳交警。电单车继续治理”转变成”深圳禁摩限电“和”支持深圳交警电单车继续治理 “。

在文本立场分析的任务, 一般目标主题包含的信息较少，而文本包含了大部分的信息。例如上面两个例子所举例的，目标主题文本分别为”Hillary Clinton“和”深圳禁摩限电 “，而 Twitter 和微博文本包含的消息较多，结合立场分析文本的特点，改善了原有的条件编码长短期记忆的网络结构，后续实验论证在多种条件编码长短期记忆的改进方案中，下面所示的网络结构具有更好的实验效果，后续在实验分析其可能的原因。

模型的具体公式如下

$$[h_1\ c_1] = LSTM^{target}(t_1, h_0, c_0) \tag{3-13}$$

...

$$[h_M\ c_M] = LSTM^{target}(t_1, h_{M-1}, c_{M-1}) \tag{3-14}$$

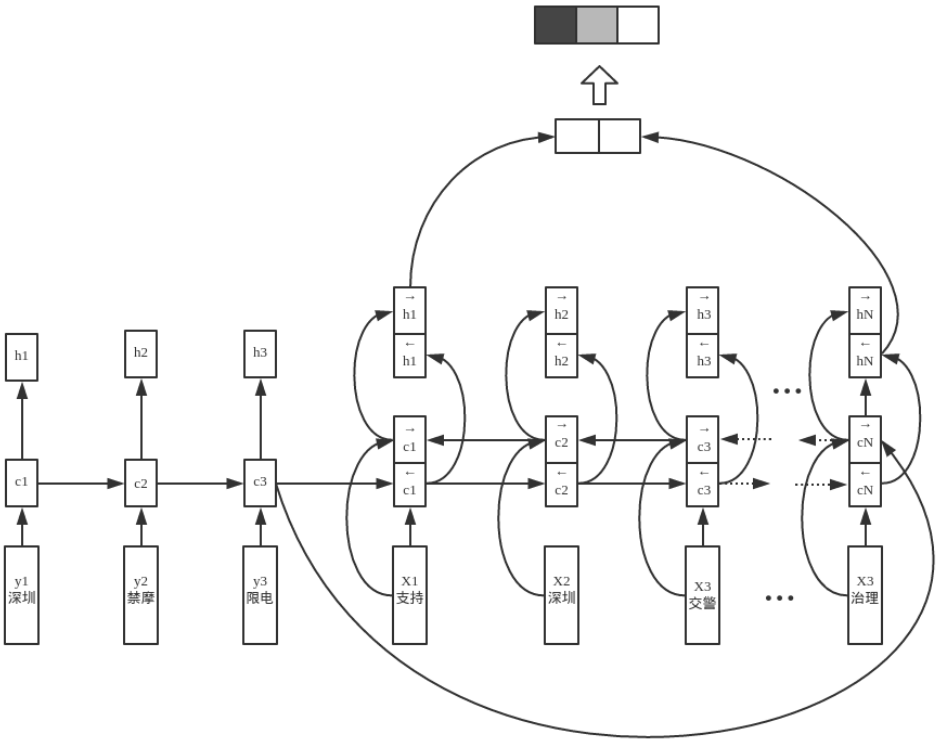

图 3-5 条件编码长短期记忆

$$[h_1^{forward}\ c_1^{forward}] = LSTM^{forward}(x_1, h_0, c_T) \tag{3-15}$$

...

$$[h_N^{forward}\ c_N^{forward}] = LSTM^{forward}(x_n, h_{N-1}^{forward}, c_{N-1}^{forward}) \tag{3-16}$$

$$[h_N^{backward}\ c_N^{backward}] = LSTM^{backward}(x_n, h_0, c_T) \tag{3-17}$$

...

$$[h_1^{backward}\ c_1^{backward}] = LSTM^{backward}(x_1, h_2^{backward}, c_2^{backward}) \tag{3-18}$$

$$c_i = Softmax(W[h_n^{forward}\ h_1^{backward}]) \tag{3-19}$$

其中 M 为主题目标的文本长度，N 为微博 〇 Twiiter 的文本长度。$LSTM$ 单向编码主题目标，$LSTM^{forward}$ 为前向编码文本信息，$LSTM^{backward}$ 为后向编码文本信息，h_0 为 LSTM 的初始化向量，C_T 为 $LSTM$ 主题目标编码的最后一个 Cell 状态，$h_1^{backward}, h_N^{forward}$ 分别为前向和后向编码的最后一个隐藏状态。

本节以“深圳禁摩限电”为话题目标，微博文本“支持深圳交警。电单车继续治理”为例, 按本节模型的 5 个层次, 描述基于条件编码长短期记忆的立场分析的过程。

(1) 输入层

先将话题目标和微博文本经过预处理操作，然后通过分词工具把话题目标和微博文本进行划分，对于同一个话题目标，微博文本分词后句子长度有可能不一致，为了方便后续神经网络框架中的批量的并行计算，通过统计选择 30 为固定长度, 长度超过固定长度进行截断操作，不够的进行补齐词表中规定 <PAD> 关键词。如例微博文本最后转换成”支持深圳交警。电单车继续治理 <PAD> ... <PAD>“

(2) 词向量嵌入层

词向量的嵌入层，此层的功能是对输入的每一次词检索其词向量 (lookup 操作)，后续实验词向量的预训练由 GloVe 模型在大量无监督语料上训练可得，预训练的词向量维度为 100，且把词向量设置为可训练，随神经网络模型的训练动态调整权重。

(3) 主题目标编码层

通过一个单向的长短期记忆 (LSTM) 模型模型编码主题目标，把分词后的主题目标“深圳禁摩限电”通过 lookup 操作取出相应的词向量，经过一个隐藏层为 64 的单向 LSTM 模型，且保留最后的记忆体的状态，如上图所示，保留 XXX 的单元的信息，给下一层文本编码做输入。

(4) 文本编码层

由于文本包含大部分信息，所以采用了双向的 LSTM 模型而不是 Rocktaschel 提出的单向的 LSTM 模型编码文本信息。如上图所示，文本编码的双向 LSTM 模型的初始 Cell 状态是由上一层的目标主题编码的最后的 Cell 状态填充，表示主题目标的消息流入到当前的双向 LSTM 模型中参与对文本的编码操作。每一个方向 LSTM 的最后的隐状态 h 作为对整个文本的最终编码表示。

(5) 全连接层

全连接层接受来自文本编码层每个方向最后的隐状态, 拼接两个隐状态信息作为最后的特征向量。全连接层的输出个数为 3，表示每个立场的预测概率。通过 softmax 激活函数归一化三个立场的概率，在预测阶段我们选取概率最大的立场当做预测的类标。

3.4 实验结果及分析

本节主要介绍条件编码长短期记忆神经网络模在 2016NLPCC 微博中文语料库和 2016SemEval 英文 Twitter 数据集中的实验结果及分析。并从实验的角度论证条件编码长短期记忆在立场分析问题上的有效性。本节包含两部分: 3.4.1 小节介绍中英文两个数据集中训练和测试文本的分布、实验评价方式以及对比方法;3.4.2 小结介绍模型训练阶段的性能调优, 以及本章实验与其他方法的比较。

3.4.1 实验数据与评价指标

（1）数据集简介

为验证条件编码长短期记忆神经网络在立场分析任务上算法的性能表现, 本节采用 2016 NLPCC 中文微博语料和 2016 SemEval 英文的 Twitter 语料两个数据集验证算法的效果。以下分别介绍两数据集的分布。

中文数据集来自 NLPCC2016 立场分析测评任务 [55] , 数据集的 5 个话题目标分别为 iPhone SE、春节放鞭炮、俄罗斯在叙利亚的反恐行动、开放二胎政策和深圳禁摩限电。所有语料都来自于新浪微博，每个微博文本的立场属于“支持”、“反对”和“其他”三者之一。NLPCC 2016 中文微博数据集的训练集、测试集按照 75% 与 25% 的比例划分, 如表 3-10 所示详细介绍每个话题目标下数据的分布。

表 3-1 训练集、测试集话题数量及立场分布比例（中文数据集）

预置话题分类	训练集数量和立场比例 (%)				测试集数量和立场比例 (%)				文本数量
	数量	支持	反对	其他	数量	支持	反对	其他	
iPhone SE	600	40.8	34.8	24.3	200	37.5	52.0	10.5	800
春节放鞭炮	600	41.7	41.7	16.7	200	44.0	47.0	9.0	800
俄在叙反恐行动	600	41.7	41.7	16.7	200	47.0	43.0	10.0	800
开放二胎政策	600	43.3	33.3	23.3	200	49.5	47.5	3.0	800
深圳禁摩限电	600	26.7	50.0	23.3	200	31.5	55.0	13.5	800
总计	3000	38.8	40.3	20.9	1000	41.9	48.9	9.2	4000

英文数据集来自 SemEval2016 Task6 stance detection[55] , 数据集的 5 个话题目标分别为 Atheism(无神论)、Climate Change is a Real Concern(气候变化真实性)、Feminist Movement(女权运动)、Hillary Clinton (希拉里克林顿) 和 Legalization of Abortion(堕胎合法化)。所有语料都来自于英文 Twitter 文本，每个 Twitter 文本的立场属于“支持”、“反对”和“其他”三者之一。不同于上述中文语料的分布，每个话

题目标英文 Twitter 语料的数量参差不齐，但总体上训练集和测试集按 70% 与 30% 的比例划分, 如表 3-2 所示详细介绍每个话题目标下数据的分布。

表 3-2 训练集、测试集话题数量及立场分布比例（英文数据集）

预置话题分类	训练集数量和立场比例 (%)				测试集数量和立场比例 (%)				文本数量
	数量	支持	反对	其他	数量	支持	反对	其他	
Atheism	513	17.9	59.3	22.8	220	14.5	72.7	12.7	733
Climate Change	395	53.7	3.8	42.5	169	72.8	6.5	20.7	564
Feminist Movement	664	31.6	49.4	19.0	285	20.4	64.2	15.4	949
Hillary Cliton	689	17.1	57.0	25.8	295	15.3	58.3	26.4	984
Legal of Abortion	653	18.5	54.4	27.1	280	16.4	67.5	16.1	933
总计	2914	25.8	47.9	26.3	1249	24.3	57.3	18.4	4163

为了和已有方法进行性能比较，本文在两数据集上都按照分别测评的比例划分出训练集和测试集。其中训练集负责模型的训练和调优, 测试集则进行最后模型的性能的评估。中英文数据集均包含 5 个不同的话题目标, 每个话题目标包含若干的话题文本，由于每个主题目标关注的内容不同且具有各自独特的语言特点。为了使模型更好的拟合每一种话题的特性, 本文首先按照中英文不同语料集合划分成两个大任务，然后根据每个语料库在细分成 5 个不同子任务，分别建立不同的条件编码长短期记忆模型。各子模型预测结束后, 统计各个子任务上的性能并汇总预测结果进行最后统一指标的计算。

（2）评价指标

中英文数据集上的社交媒体文本的立场结果有“支持”，“反对”，“其他”，可以把任务当成一个三分类任务，但是由于三个类在不同的数据集下的不同主题目标的分布有可能很不平均，如果只单独用正确率（Accuracy）作为评测指标则缺失了评价指标的客观性。本文 c 采取了两个评测任务都使用的”支持“和”反对“的 F1 指标的微平均（micro-average）作为最后模型的评测指标，但是为了更清晰的评价每个主题目标的性能，每个目标主题也会单独计算微平均评测指标。为了清晰地解释指标的含义，列举以下公式说明

首先定义准确率 (Precision, P)、召回率 (Recall, R), 如公式 3-20 和公式 3-21 所示。

$$P = \frac{TP}{TP + FP} \tag{3-20}$$

$$F = \frac{TP}{TP + FN} \tag{3-21}$$

TP：正样例预测为正样例的个数

FP：负样例预测为正样例的个数

FN：正样例预测为负样例的个数。

精确率计算的是所有” 正确被检索的样例 (TP)” 占所有” 实际被检索到的 (TP+FP) 样例的比例。召回率计算的是所有” 正确被检索的运力 (TP)” 占所有” 应该检索到的样例 (TP+FN)” 的比例。如果要同时考虑精确率和召回率，则需要采样两者的调和平均值，也称为 F1 值，其定义如公式 3-22 所示

$$F1 = \frac{2PR}{P+R} = \frac{2TP}{2TP+FP+FN} \tag{3-22}$$

有关”支持“立场 F1 值的计算，“支持”类标作为正样本，“反对”后“其他”作为负样本。因此其计算公式如下所示

$$F1_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor}+R_{favor}} \tag{3-23}$$

同样的”反对“立场 F1 值的计算，“反对”类标作为正样本，“支持”后“其他”作为负样本。因此其计算公式如下所示

$$F1_{against} = \frac{2P_{against}R_{against}}{P_{against}+R_{against}} \tag{3-24}$$

立场分析总的平均指标 F1 的微平均（Micro-average）的计算公式如下

$$F1_{average} = \frac{F1_{favor}+F1_{against}}{2} \tag{3-25}$$

3.4.2 实验数据预处理与模型参数设计

本文实验的社交媒体立场分析数据都来自于新浪微博或者 Twitter, 此类网络语言文本具有文本形式极度不规范，口语化严重。Twitter 与微博是由网络用户即兴创作的短文本, 用户在用词、语法等方面随意性较大, 文本形式与新闻、维基百科等语料具有很大的差异。由于 Twitter 和微博一般有对单条文本长度做出文本长度的限制, 因此这些网络文本中常含有缩写、俗语、流行词汇等元素, 同样也存在语法成分缺失的问题。除此之外, 文本中大量存在的网页链接、“@ 某用户”和“# 某话题”等功能标记。

由于社交媒体文本根据上述特点，分别对中英文语句进行预处理操作。删除语料中其中的大量 URL 信息，由于“# 某话题”等话题标签对立场分析有很重要的影响，因此把话题标签消息保留了下来，对于英文所有词语全转换成小写拼写，分

词工具采用了 CMU 专门为 Twitter 开发的 Twitter NLP tool 里面的分词模型 [XXX], 中文的分词工具采用较为稳定的结巴分词工具。对于网络社交文本分词后可转换成一些词语的序列信息，对于在训练集中出现次数小于 2 词的词语归为低频词，为了减少模型的参数，加大模型的泛化能力，把所以的低频词转换成一个统一的词汇用“UNKNOW”标识。同时为了兼容现在主流的基于批量更新的深度学习框架，对文本的长度进行固定操作。英文的文本固定为 30 个词的长度，中文微博文本固定为 50 的词长度。

英文语料的词向量训练通过 GloVe 算法在 20 亿 Twitter 文本语料训练而成，其中包含了 120 万个词，词向量维度为 200。中文微博语料是通过 Word2Vec 算法在大量微博语料中训练所得，其中包含了 17 万个词，词向量的维度也为 200。

本文建立了针对 5 个不同话题目标的子模型, 使用相同的条件编码长短期记忆模型参数。取出训练集合的 10% 作为验证集参与模型的选择。通过实验发现，当对主题目标和文本编码 LSTM 模型的隐藏层单元设置为 64，embedding 层的 dropout 的概率设置为 0.2，当对主题目标和文本编码 LSTM 模型内部记忆的 dropout 概率设置为 0.3，每次以 32 个样本作为 mini-batch 参加参数更新，选取 0.001 的学习率的 Adam 优化方法。

表 3-3 基于条件双向编码长短期记忆的立场分析实验超参数集

序号	超参数名称	数值
1	LSTM 隐藏层单元	64
2	Embedding 层 dropout	0.2
3	LSTM 内部 drought	0.3
4	批处理大小	32
5	L2 正则化参数	1e-6
6	全量迭代次数	50
7	梯度优化方法	Adam
8	学习率	0.001

3.4.3 模型对比实验结果及分析

通过网格搜索调节不同参数，使得基于条件双向编码长短期记忆模型在各个数据集上取得最优的效果。调优过程中，在模型的词向量抽取层、LSTM 内部记忆单元层、和最后的特征向量层加入适当的 dropout 可以使得模型更加的稳定，Dropout 一定层度上减少了模型过拟合的趋势，增加了模型的泛化能力，通过调节最优参

数，模型在 SemEval2016 数据集上的表现表 3-4 所示。

表 3-4 条件双向编码长短期记忆模型在各个话题试验性能（SemEval 数据集）

主题目标	P_{favor}	R_{favor}	F_{favor}	$P_{against}$	$R_{against}$	$F_{against}$	$F_{average}$
Atheism	0.3824	0.4063	0.3939	0.8323	0.8375	0.8349	0.6144
Climate Change	0.7389	0.9431	0.8286	0.0000	0.0000	0.0000	0.4143
Feminist Movement	0.3924	0.5345	0.4526	0.7205	0.6339	0.6744	0.5635
Hillary Clinton	0.6316	0.2667	0.3750	0.6638	0.8837	0.7581	0.5666
Legal. of Abortion	0.4630	0.5435	0.5000	0.7770	0.6085	0.6825	0.5912
合并统计	0.5743	0.6480	0.6112	0.7396	0.7231	0.7314	0.6713

如标 3-4 所示，以下分析模型的结果。条件双向编码 LSTM 模型在所有类标的 micro-F1 值的指标为 0.6713，其总结所有主题目标“支持”立场的 F1 值为 0.5789，而“反对”立场的总体 F1 值有 0.7418。从这个数据可以看出来模型对反对立场有更好的性能，模型会有这样的偏置的原因是由于原英文数据的分布有关，原英文数据集的分布如表 3-2 所示，训练集 0.47 的样本是持”反对“立场，测试集合也有 0.57 的持反对立场。而“支持”立场分别只占了 0.25 和 0.24。由于有更多的“反对”样本，条件双向编码 LSTM 模型能更好的识别”反对“立场中的分类模型，并且模型在不平均数据集上会对高频的类有先天的偏置。这两点导致了模型在“反对”立场上的性能比“支持”有更好的性能。单独分析具体话题目标，“Climate Change is a Real Concern”主题目标的 F1 值只有 0.41，远低于其他主题目标的指标。从原数据分析发现，“Climate Change is a Real Concern”主题目标下”反对“立场只占原数据集的 0.03，”支持“的占有率为 0.53，”支持“的样本是“反对”的 10 多倍了。而从模型最后对于“反对”测试集准确率为 0，召回率也为 0。可以看出最后模型都未预测任何一个样本为“反对”立场，这导致了模型在“Climate Change is a Real Concern”的准确率和召回率都为 0，其次训练集和测试集在“Climate Change is a Real Concern”上的分布也有较大的差距，支持立场在训练和测试集合中有将近 20% 的差距，这样一点程度上影响模型在此主题目标下的性能，最终影响了整体的模型性能。

如标 3-4 所示，整体模型的 Micro-F1 值为 0.6713。远高于性能好的子主题目标”Atheism” 上的 F1 值 0.6114。其中的原因有可能是各个主题目标的“反对”和“支持”分布较不均匀，例如“Climate Change is a Real Concern”中“支持”立场占了 0.53，而“Hillary Clinton”中”反对“立场占据了 0.57。相对来说单个主题目标中“支持”与”反对“的分布很均衡，但整体综合 5 个主题目标后，“支持”与”

反对“却变得相互平衡了。因此导致”支持”和“反对”立场都拥有了较好了 F1 指标，进一步 micro-F1 的指标也会优于各个主题目标的 F1 指标。

在 NLPCC2016 数据集上，采取对编码主题目标和微博文本的 LSTM 隐藏单元为设为 100，其他超参数和 SemEval 英文数据一样。在此数据集上综合 5 个主题目标，“支持”立场的 F1 值为 0.665，“反对”立场的 F1 值为 0.707，总体的 Micro-F1 值为 0.686。

在 NLPCC2016 数据集上，每个主题目标的性能如表 3-5 所示，只有“俄罗斯在叙利亚的反恐行动”的平均 F1 指标只有 0.583，其他指标都上了 0.6，而“春节放鞭炮”主题目标的 F1 平均值到达了 0.745，说明模型在此数据集的性能相对较好。原始数据的分布也造成此现象的根本原因。中文文本的训练数据和测试数据如表 3-10 所示，除了“深圳禁摩限电”主题目标外，其实 4 个主题目标的”支持“立场大概占据 0.4 左右的比例，而“反对”立场则占据 0.35 左右的。而且测试集合和训练集合的分布相似度极高，因此模型在中文数据集相对的性能就比较高。

表 3-5 基于子话题分别训练的条件双向编码长短期记忆模型试验性能（SemEval 数据集）

主题目标	P_{favor}	R_{favor}	F_{favor}	$P_{against}$	$R_{against}$	$F_{against}$	$F_{average}$
iPhone SE	0.5844	0.6000	0.5921	0.6842	0.6250	0.6533	0.6227
春节放鞭炮	0.6792	0.8182	0.7423	0.8049	0.7021	0.7500	0.7461
俄在叙反恐行动	0.6267	0.5000	0.5562	0.5299	0.7209	0.6108	0.5835
开放二胎政策	0.6694	0.8384	0.7444	0.8413	0.5579	0.6709	0.7076
深圳禁摩限电	0.6308	0.6508	0.6406	0.7674	0.9000	0.8285	0.7345
总计	0.6443	0.6874	0.6658	0.7099	0.7055	0.7077	0.6868

在文本立场分析中，主题目标信息对文本立场分析有较大的影响。首先在文本立场分析的定义上，任何文本在其主题目标未定义清楚前，其文本立场是无法决策的。单独的文本其本身只有情感分析而无立场分析，因此本文以此为出发点，为论证在立场分析中，合理利用主题目标的信息能提高立场分析的性能，设计了以下模型。

用单向 LSTM 直接编码文本 Text 信息，不引入主题目标信息，以下简称 **Text-Only**。

用两个不同 LSTM 模型分别编码主题目标 Target 和文本 Text 信息，两个 LSTM 独立编码各自的信息，最后拼接两者的编码向量作为最后参与分类的特征向量，以下简称 **Text-Target**。

用两个不同 LSTM 模型分别编码主题目标 Target 和文本 Text 信息，对主题目

标 Target 最后 LSTM 模型的记忆单元 (Cell) 状态做为对文本编码 LSTM 模型的初始状态，构成以主题目标 Target 为条件的 Text 文本编码，取 Text 编码 LSTM 的最后一个隐状态作为最后的特征向量，以下简称 **Text-on-Target**。

针对立场分析数据的特点，主题目标所包含的信息相对有限，而文本包含绝大部分信息的特点。改进了条件编码的模型，与（3）相同用单向的 LSTM 编码主题目标信息，为了充分提取文本的信息，采用双向 LSTM 模型来提取文本的信息，类似于 (3) 的做法，编码文本的双向 LSTM 模型的记忆单元来自于对主题目标编码的 LSTM 的最后记忆单元状态，以下简称 **BiText-on-Target**。

首先为了初步验证主题目标的信息对立场分析是否具有提升作用，通过对比 Text-Only 和 Text-Target 模型的性能可以初步得出主题目标的信息对立场分析是否有促进作用，然后为了进一步说明条件编码 LSTM 模型是否能跟好的利用主题目标的消息参与对文本的编码，设计了 Text-Target 模型和 Text-on-Target 模型的对比实验。最后为了对比我们改进过了的条件编码 BiText-on-Target 模型是否能更好编码文本信息，设计了 Text-on-Target 模型和 BiText-on-Target 模型的对比实验。各模型的方式建立基础如表 3-6 所示。

表 3-6 模型比较

主题目标	文本编码方式	是否引入主题目标	引入主题目标方式
Text-Only	单向 LSTM	否	无
Text-Target	单向 LSTM	是	直接拼接
Text-on-Target	单向 LSTM	是	条件编码
BiText-on-Target	双向 LSTM	是	条件编码

对比实验各个模型在两个数据集分别调整超参数，在个模型中做多组的实验取其最优结果当做模型的预测结果。为说明各个模型在具体各个主题目标的性能，以下以英文 SemEval 数据集为例，计算各模型在不同主题目标下的性能，各模型在英文数据集各主题目标下性能如图 3-6 所示，除了在”Climate Change is a Real Concern“主题目标下各模型的性能差距相对较少，具体原因由此主题目标信息的原始信息分布有关，上文已经做过相应的说明，这里就不在赘述。而其他主题目标下模型的性能还是有比较明显的差距的。

从图 3-6 可知只用文本信息的 Text-Only 模型在所以主题目标的效果都不是很好，只有在”Hillary Clinton“主题目标上能稍微领先结合主题目标信息的 Text-Target 模型。Text-Target 模型在所有主题上比 Text-Only 模型无明显的提升，在”Hillary Clinton“主题目标上的指标还低于 Text-Only 模型，这样侧面说明两个独立

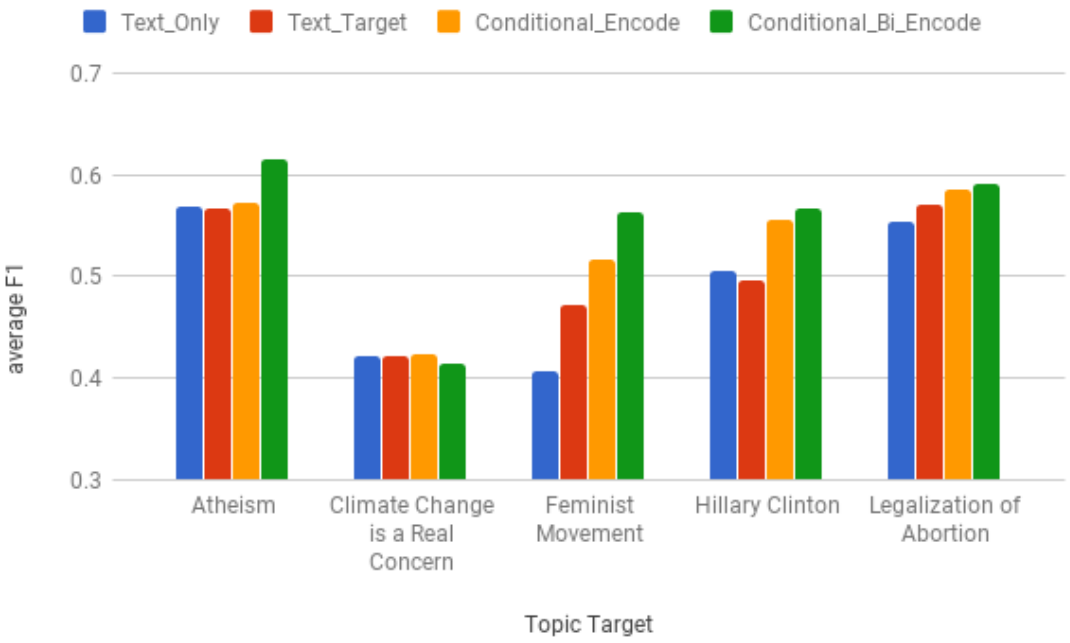

图 3-6 模型各主题目标性能分析

的 LSTM 模型编码主题目标信息和文本信息的模型比单独编码文本信息模型无显著的提高，独立的编码模型无法发现主题目标和文本信息相互之间作用的模式。

利用条件编码的 Text-On-Target 模型在各主题目标的性能相对与 Text-Only 和 Text-Target 模型都有都有显著的提升，这证明了条件编码模型能较好的结合主题目标信息和文本信息从而能找到立场分析的模式，其中的理论依据在于编码文本信息的 LSTM 模型是以编码主题目标信息 LSTM 为条件，信息从编码主题目标 LSTM 模型流入至编码文本信息的 LSTM，在未编码文本信息前就有相应的”先验知识“，对文本信息编码的 LSTM 模型根据主题目标的”先验知识“，能更好抓取文本中有关立场分析的模式信息。

结合社交媒体立场分析中主题目标的信息有限和文本信息较为丰富的特点，将条件编码中对文本的编码转换成双向的 LSTM 模型，已达到更好编码文本信息的目的。在 Semeval 英文数据集上也论证了这一想法。基于上述出发点改进过后的 Bi-Text-On-Target 模型在除”Climate Change is a Real Concern“立场以外的其他所以主题目标的性能都优于其他三个模型。

汇总各个主题的指标，各模型的总体”支持“与”反对”立场的 F1 值与 Micro-F1 值如表 3-8 所示。

其中两个数据集中 Text-Only 模型取得 0.644 和 0.678 的微平均 F1 值，而 Text-

表 3-7 模型整体性能对比（SemEval 数据集）

模型	P_{favor}	R_{favor}	F_{favor}	$P_{against}$	$R_{against}$	$F_{against}$	$F_{average}$
Text-Only	0.5364	0.5822	0.5593	0.7258	0.7329	0.7293	0.6443
Text-Target	0.4564	0.6711	0.5637	0.7594	0.6224	0.6909	0.6273
Text-on-Target	0.5776	0.6118	0.5947	0.7234	0.7608	0.7421	0.6684
BiText-on-Target	0.5743	0.6480	0.6112	0.7396	0.7231	0.7314	0.6713

表 3-8 模型整体性能对比（NLPCC 数据集）

模型	P_{favor}	R_{favor}	F_{favor}	$P_{against}$	$R_{against}$	$F_{against}$	$F_{average}$
Text-Only	0.6465	0.6635	0.6550	0.7275	0.6769	0.7022	0.6786
Text-Target	0.6297	0.6778	0.6538	0.7333	0.6299	0.6816	0.6677
Text-on-Target	0.6504	0.7017	0.6761	0.7450	0.6871	0.7161	0.6961
BiText-on-Target	0.6721	0.6850	0.6785	0.7146	0.7219	0.7182	0.6984

Target 模型虽然在各个主题目标下的 F1 值接近 Text-Only 的值 F1 值，但是在总体上的微平均值只有 0.627 和 0.667，分别少了 Text-Only 模型 1 个多点，也一点程度上说明如果主题目标信息引入的方法不对（直接拼接向量）后，有可能导致模型的实际泛化能力还下降了，导致整体的效果还不如子考虑文本信息本身，因此在文本立场分析任务中，以合适的方式引入主题目标建立合适的模型是挺重要的一个环节。

以条件编码结合主题目标和文本的模型在两个数据集上都取得较好的结果，其中 Text-On-Target 模型在中英数据集上取得 0.668 和 0.698 的平均 F1 值，比 Text-Only 模型在两个模型中分别高了 0.024 和 0.020，说明引入主题目标作为对文本编码的”先验知识“能显著的提高模型在文本立场分析的性能。Text-on-Target 比 Text-Target 模型在两个数据集上高 0.044 和 0.031 的微 F1 值，此处提升的比对比 Text-Only 模型还高，说明以条件编码结合主题目标信息比直接单独拼接特征向量的方法更能挖掘文本和主题目标之间的立场关系。

对比 Text-On-Target 模型和 BiText-On-Target 模型在各数据集上的表现，BiText-On-Target 在英文数据集上取得 0.6712 的微平均 F1 值，Text-On-Target 模型取得了 0.668 的的 F1 值。BiText-On-Target 在英文数据集上比 Text-On-Target 高出 0.003 的微平均 F1 值，因此在英文数据上双向 LSTM 模型编码文本信息比单向 LSTM 模型更好，这点由于双向 LSTM 有更好的提取文体特征的能力。但在中文数据集合上，Bi-Text-On-Target 比 Text-On-Target 模型高出 0.002。双向的 LSTM 模型不管对中文还是英文的文本的建模能力更强。基于主题目标信息条件双向的 LSTM 编码

模型在四个模型中的性能最高。

上述实验论证和合适的方式引入主题目标信息可以提升文本立场分析的性能，为了验证提出的条件编码 LSTM 模型的有效性和不足，以下通过实验结果对比其他研究人员在文本研究中模型。由于两个数据集来源于不同的评测任务，因此分别从 SemEval2016 英文数据集和 NLPCC2016 中文数据集挑选出较好系统和本章提出的条件编码 LSTM 进行比较。

在 SemEval 数据集合上，引入以下模型是作为对比模型

（1）MITRE, SemEval2016 Twitter 立场分析评测任务第一名，收集大量的无标注数据。通过分析筛选出多个 # 话题标签，建立 LSTM 模型去预测文本的标签。先训练预测标签的模型，而后在预训练好的模型上做立场分析的任务上的微调，此模型基于大量无标注样本的迁移学习，需要大量无标注任务和手工筛选特别话题标签。

（2）TakeLab SemEval2016 Twitter 立场分析评测参赛模型，融合了随机森林，逻辑回归，支持向量机等多种机器学习的集成学习方法，各模型利用大量人工构造的特征。

（3）ECNU SemEval2016 Twitter 立场分析评测华东师范大学参赛模型，采用了传统语义特征、主题目标特征、相似性特征、情感字典特征、主题模型特征和词向量特征的特征工程的传统机器学习方法

（4）BiText-On-Target 本章出的基于主题目标的条件的双向 LSTM 文本编码模型

四个模型在 SemeEval2016 的各主题目标下单个平均 F1 值如图 3-7 所示 所以模型在”Climate Change is a Real Concern“主题目标上的性能都较差，没有一个模型的性能超过了 0.43 的平均 F1 值。数据类标分布极度不平均，训练集与测试集分布差异大的问题的问题影响了所有模型，4 个模型都没克服这个缺陷。基于特征工程的 TakeLab 和 ECNU 模型对”Legal of Abortion“主题目标的性能相对较好，基于深度学习的 MITRE 模型和 BiText-On-Target 模型相对较差，在此主题目标下，基于统计的特征具有更好的立场分析区分度。TakeLab 在”Atheism“和”Hillary Clinton“主题目标的效果远高于其他模型，原因是其运用了集成随机森林，逻辑回归，支持向量机等多种分类器，模型相对强健。BiText-On-Target 模型在所有的主题目标下都拥有相对稳定的效果，虽然没在任意主题目标取得最好的立场分析效果，但所有主题目标的效果都不是很差，相对于其他三个模型更加稳定和强健。

上述模型在 SemEval 数据集下的综合”支持“和”反对“的 F1 值和总体微平均 F1 值的性能如下表 3-8 所示

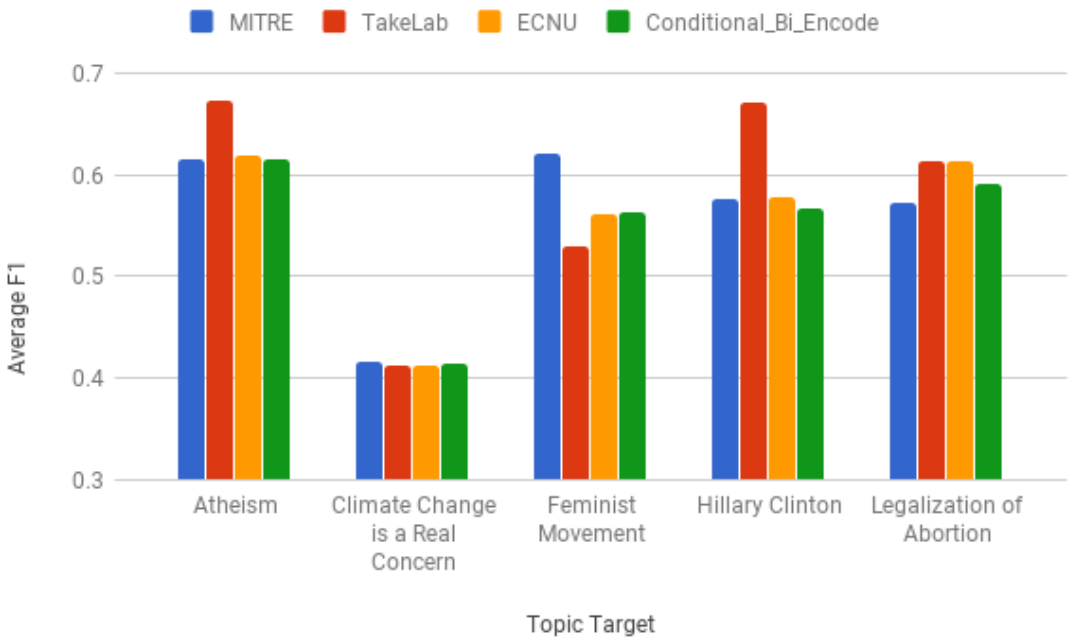

图 3-7 模型在 SemeEval2016 各主题目标平均 F1 值分布

评测最好的模型 MITRE 的的微平均 F1 值为 0.678，本章提出的模型 BiText-On-Target 取得了 0.671 的平均 F1 值，此模型在评测中取得 3（20）的成绩。虽然模型的性能比 MITRE 的相比低了 0.007，但 MITRE 需要大量的无标注数据，TakeLab 模型虽然在某些主题目标下取得较好的 F1 平均值，但是对于总体的”支持“和“反对”立场微平均 F1 值的效果不如 BiText-On-Target 模型。

表 3-9 各模型在立场分析测试集中的性能表现（SemEval 数据集）

主题目标	F_{favor}	$F_{against}$	$F_{average}$
MITRE	0.5932	0.7633	0.67820
TakeLab	0.6093	0.7273	0.6683
ECNU	0.6055	0.7054	0.6555
BiText-On-Target	0.6112	0.7314	0.6713

3.4.4 模型对比实验结果及分析

表 3-10 基于子话题分别训练的条件双向编码长短期记忆模型试验性能（NLPCC 数据集）

主题目标	P_{favor}	R_{favor}	F_{favor}	$P_{against}$	$R_{against}$	$F_{against}$	$F_{average}$
iPhone SE	600	40.8	34.8	24.3	200	37.5	52.0
春节放鞭炮	600	41.7	41.7	16.7	200	44.0	47.0
俄在叙反恐行动	600	41.7	41.7	16.7	200	47.0	43.0
开放二胎政策	600	43.3	33.3	23.3	200	49.5	47.5
深圳禁摩限电	600	26.7	50.0	23.3	200	31.5	55.0
总计	3000	38.8	40.3	20.9	1000	41.9	48.9

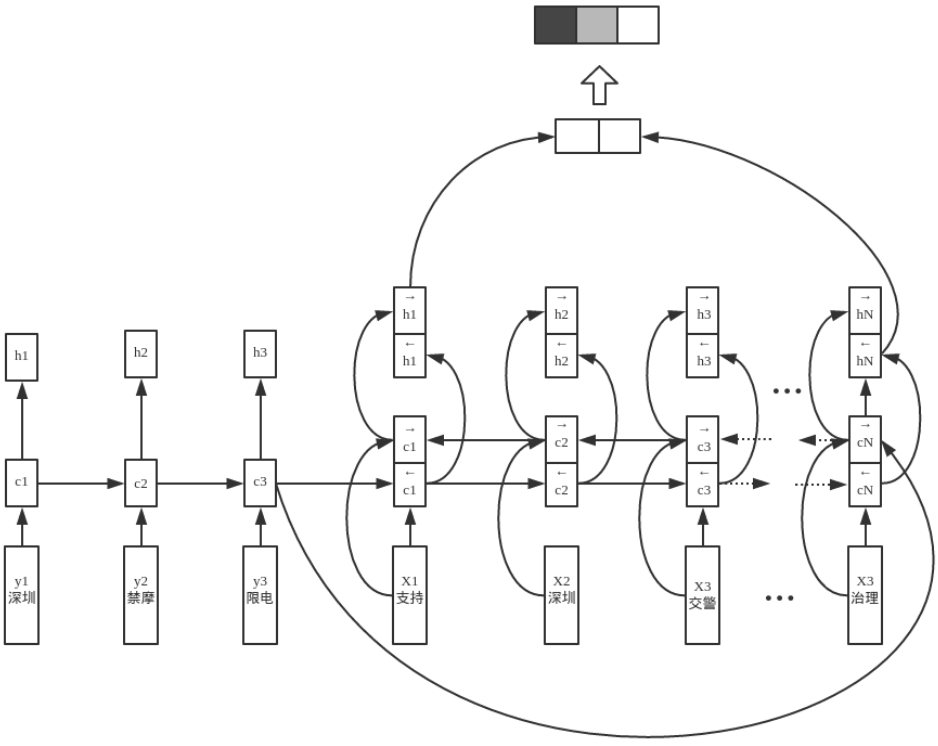

图 3-8 条件编码长短期记忆

第 4 章 基于注意力机制的卷积神经网络社交媒体文本立场分析

4.1 引言

近年来，深度学习的研究越来越深入，在各个领域也都获得了不少突破性的进展。基于注意力（attention）机制的神经网络成为了最近神经网络研究的一个热点，引起了研究者的广泛关注。在神经网络实现预测任务时，引入注意力机制能使训练重点集中在输入数据的相关部分，而不是无关部分。在社交媒体文本立场分析任务中, 不管是基于传统特征工程的机器学习模型 SVM 和基于端到端的深度学习的 RNNs，CNNs 等神经网络模型，都忽略主题目标信息对社交媒体立场分析的重要作用，这样建立模型存在的不合理性，为了更好的引入主题目标信息帮助提高社交媒体文本立场分析任务的性能，在本章引入基于注意力机制的神经网络模型。通过不同的权重，注意力机制能根据主题目标更好的对微博或者文本信息进行不同重要性的关注，使模型更加聚焦在对主题目标立场分析重要的信息上。实验结果表明, 基于注意力机制的神经网络模型在社交媒体文本立场分析任务中取得了相当不错的效果, 证明了注意力机制的有效性。

本章 4.2 节首先介绍深层记忆网络的重要机制及架构, 4.3 节介绍深层记忆网络在社交媒体文本立场分析中的应用,4.4 节介绍实验设置和结果的分析,4.5 节作出本章小结。

4.2 注意力机制的卷积神经网络模型的社交媒体文本立场分析

近年来，基于深度学习已经在图像识别、语音识别和自然语音处理上上获得了重要的进展。在自然语言处理任务中，注意力机制在神经网络机器翻译、序列标注、层次性文本分类上取得突破性提高。注意力机制将对信息呈现不同的关注程度，通过聚焦在重要的信息上，达到模型性能的特点。本节!!!!!!

4.2.1 注意力机制

注意力机制最早应用在图像识别领域上，研究人员研究的动机其实也是受到人类注意力机制的启发。人们在进行观察图像的时候，并不是一次就把整幅图像

的每个位置像素都看过，大多是根据需求将注意力集中到图像的特定部分。而且人类会根据之前观察的图像学习到未来要观察图像注意力应该集中的位置。

注意力机制除在图像识别和语音识别上取得巨大的成功，近期在基于 RNNs 的端到端的编码解码的机器翻译、序列标注和层次文本分类等任务取得突破。在自然语言处理任务中，首先引入注意力机制的是神经网络机器翻译。神经网络机器翻译其实就是一个典型的序列到序列模型，也就是一个编码解码模型。由于基于注意力机制的卷积神经网络模型的注意力机制很类似于机器翻译中的注意力机制，这里将简单介绍下机器翻译中的编码解码注意力机制模型。

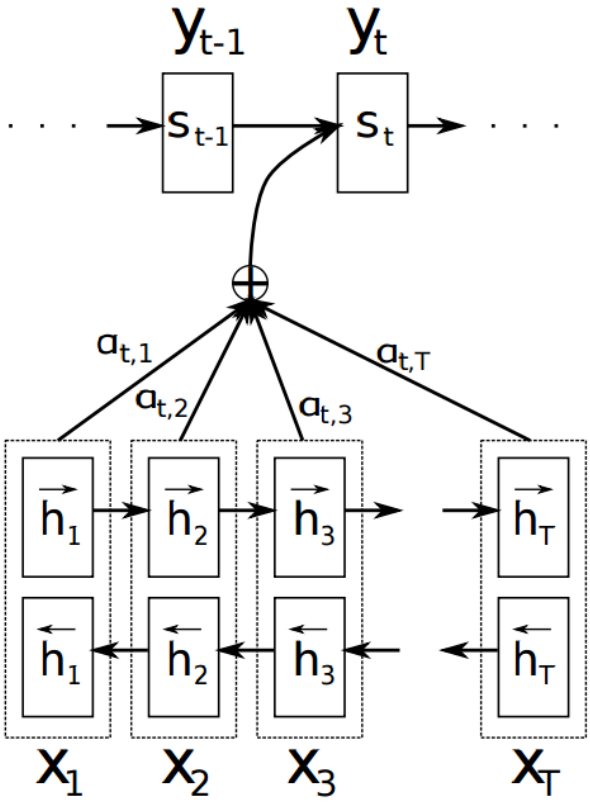

图 4-1 注意力机制神经翻译模型

传统的神经翻译模型编码阶段里面用的是 RNNs 模型，这样每个单词的表达不仅能包含前一个单词的信息，还可以包含后一个; RNNs 按输入序列的顺序，生成同样顺序的隐藏层状态，这样它就既包含当前单词的前一个单词信息，也包含后一个信息; 这个状态之后将被用于解码阶段部分。

$$h_t = f(x_t, h_{t-1}) \tag{4-1}$$

其中编码信息为

$$c = q(\{h_1, ...h_{T_x}\}) \tag{4-2}$$

其中 h_t 是时间 t 的隐藏状态，c 向量是从序列信息得到的压缩向量，通常 f 和 q 是非线性函数，例如 SutSkever[XXX] 等用的是 LSTM 作为 f 函数，把前向循环网络的最后一个隐状态当做最好的压缩向量，$q(\{h_1, ...h_{Tx}\}) = h_T$

在解码阶段，解码器的作用是根据编码的信息和已经输出的信息来预测下一个词最有可能的单词，可以用下面公式 4-3 表示

$$p(y) = \prod_{t=1}^{T} p(y_t|\{y1, ..., y_{t-1}\}, c) \tag{4-3}$$

当 $y = (y1, ..., y_T)$, 若用 RNN 式的结构，每个的条件概率如公式4-4

$$p(y_t|\{y1, ..., y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \tag{4-4}$$

而基于注意力机制的神经翻译模型的条件概率公式如下

$$p(y_t|\{y1, ..., y_{t-1}, x\}, c) = g(y_{t-1}, s_t, c_i) \tag{4-5}$$

其中 s_i 是 RNN 的隐状态，其中的计算公式如下

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{4-6}$$

其中从传统的神经翻译模型做出的改进是对于每一个不同的输出，它所依据的上下文压缩向量 c_i 不在是同一个压缩向量 c,c_i 的计算是通过 $(h_1, ..., h_{T_x})$ 的加权所得。h_i 虽然包含所以所有的输入信息，但是主要存储了 x_i 的信息。其中 c_i 的计算公式如4-7所示

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_i \tag{4-7}$$

其中注意力变量 αij 的计算公式如下

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{4-8}$$

其中 e_{ij} 的计算公式如下

$$\alpha_{ij} = a(s_{i-1}, h_j) \tag{4-9}$$

注意力机制的加入使得神经网络翻译模型具有更强的对句子翻译建模能力，使模型在翻译当前词语时关注和词语的更相关的内容，忽略不相关的内容。通过可视化注意力 αij 可显示看出在翻译不同词语时，对原语言关注的内容是不一样的，例如在中文翻译中的注意力矩阵如下图4-2所示

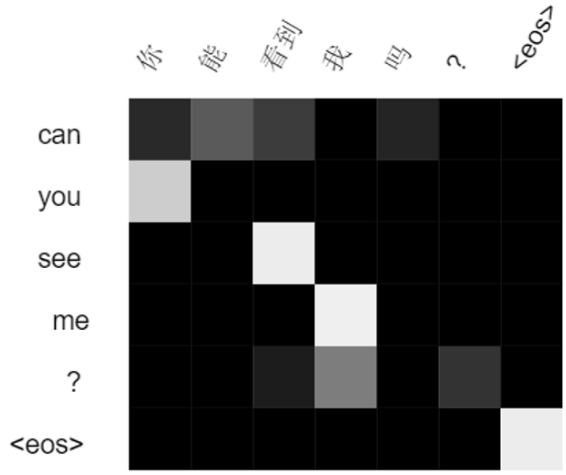

图 4-2 中英文翻译注意力矩阵

4.2.2 卷积神经网络

卷积原是常用在积分变换等领域的数学概念。现在被广泛应用于计算机视觉、自然语言处理和语音识别等任务中。卷积神经网络是对神经网络模型的改进。卷积层把原来的全连接改成局部连接权值共享，既可以大大缩减参数个数，又可以适应不同区域统计特性一样的特征。下采样层可以利用局部信息采样特征图得到更低纬度的特征，提高模型的泛化能力，防止训练的分类器过拟合。

根据不同任务，卷积神经网络的输入矩阵处理方式也不全相同。用于图片分类时，输入矩阵为像素点矩阵，用于文本分类时，输入矩阵为相同维度的词向量以一行一个词拼接。卷积神经网络使用预设石村的卷积核作为过滤器，与输入矩阵做卷积操作。

以文本卷积神经网络为例，当输入矩阵的词向量维数为 d, 句子长度为 l, 输入矩阵的大小为 $A \in R^{l \times d}$, $A[i:j]$ 表示输入矩阵中从第 i 行到第 j 行的子矩阵。尺寸为 h 的过滤器对矩阵 A 做卷积操作，输出矩阵 $O \in R^{s-h+1}$ 如公式 4-28 所示，

$$O_i = W \cdot A[i:i+h-1] \tag{4-10}$$

其中，i 为从 1 到 $s-h+1$ 的数列。

将输出矩阵与偏置 b 相加，放入激活函数中处理，即可得到卷积层的输出 C_i。

卷积层一般会连接下采样层。在自然语言处理任务重，下采样层通常使用 KMAX 池化函数，所以下采样层也被称为池化层。池化层既可以对卷积层的输出再次降维从而组合更大窗口范围的特征，同时也可以将不同尺寸的卷积核得到的特征构建为固定长度的输出，避免了不同输入矩阵大小表示的差异。之后的向

量通常会连接 Dropout 层，随机抛弃一些节点，也是为了减少过拟合现象同时可以增强网络的训练效果。最后的全连接分类层通常使用 softmax 函数，基于多卷积核文本的 CNN 分类框架图如 4-3 所示

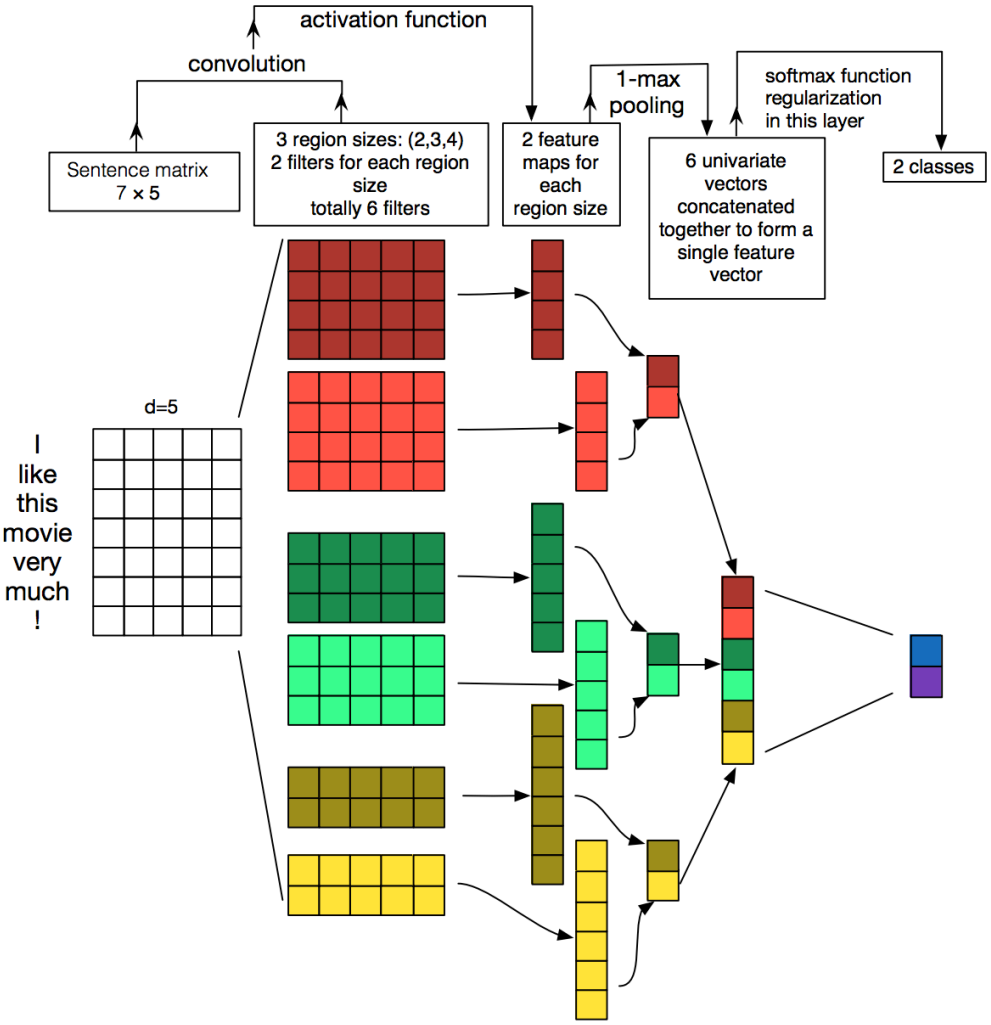

图 4-3 多卷积核文本 CNN

4.2.3 注意力机制卷积神经网络社交媒体文本立场分析

基于注意力机制在神经网络翻译模型、序列标注、层次文本分类等领域取得的成功。文本多卷积核 CNN 模型在文本分类上的也取得了较好的效果。在社交媒体的文本立场分析上，由于当前研究没有一种有效把主题目标信息以合适的方式结合在文本的立场分析中，但是主题目标对文本立场分析又有着举足轻重的影响。为了主题目标信息没有合适引入的缺点，本文提出一种以注意力机制引入主题目

标信息，在注意力的基础上，用多卷积核文本 CNN 模型提取文本立场的模型，本节结合立场分析数据集中的中文实例数据, 描述注意力机制卷积神经网络在该问题中的主要工作流程和重要计算方法。 文本引入了双向 GRU（Gated recurrent unit）

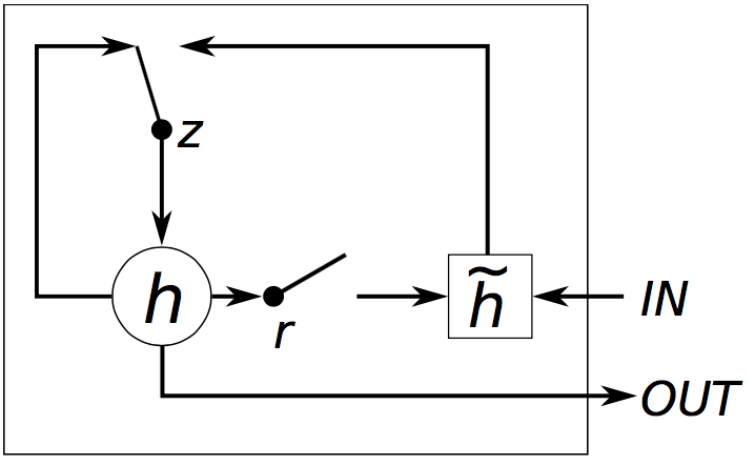

图 4-4 GRU 内部结构

用来编码文本信息，相对于 RNN 单元，GRU 也能够缓解递归神经网络的梯度消失的问题，相对于 LSTM 单元，GRU 具有更少的参数和更快的运算速度，因此本文利用例如 GRU 编码信息。相对于 LSTM 单元有输入门、遗忘门和输出门的三个门单元，按图 4-4 所示,GRU 采取了一种更加简洁的建模方式，只需要保持更新门和重置门，更新门表示隐藏单元有多少信息需要保留下来，重置门表示都是隐藏单元信息参与输入的更新，同时也删除了 LSTM 的记忆单元，具体的门的更新和输出更新公式如下

$$z_t = \alpha_g(W_z x_t + U_z h_{t-1} + b_z) \tag{4-11}$$

$$r_t = \alpha_g(W_r x_t + U_r h_{t-1} + b_r) \tag{4-12}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \alpha(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{4-13}$$

本节以“深圳禁摩限电”为话题目标，微博文本“支持深圳交警。电单车继续治理”为例, 按本节模型的 5 个层次, 描述注意力机制卷积神经网络的立场分析的过程，具体模型的结构如图 4-5 所示

（1）输入层

先将话题目标和微博文本经过预处理操作, 然后通过分词工具把话题目标和微

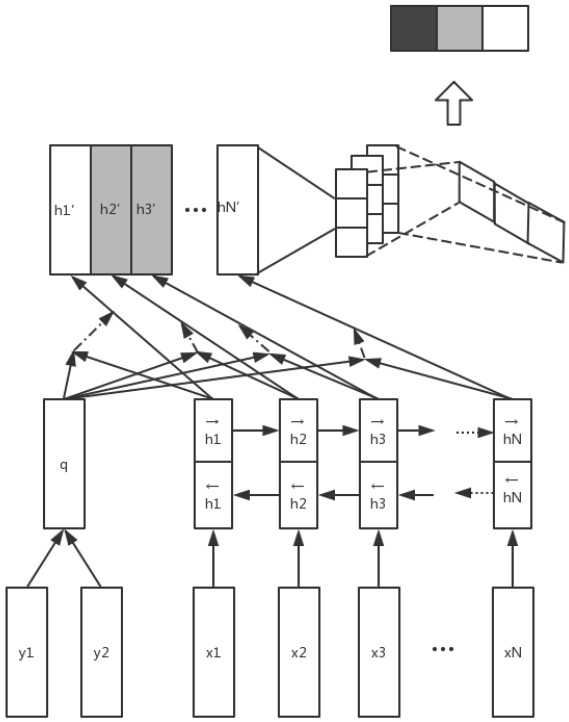

图 4-5 条件编码长短期记忆

博文本进行划分, 对于同一个话题目标, 微博文本分词后句子长度有可能不一致, 为了方便后续神经网络框架中的批量的并行计算, 英文语料统计选择 30 为固定长度, 中文语料统计后截取 50 为固定长度，长度超过固定长度进行截断操作, 不够的进行补齐词表中规定 <PAD> 关键词。如上述微博文本最后转换成”支持深圳交警。电单车继续治理 <PAD> ... <PAD> “，而话题目标也需要进行分词操作，例如例子中“深圳禁摩限电”根据结巴分词分割成“深圳”、“禁摩”和“限电”三个词组。经过此层后主题目标信息和微博文本信息可以转换成以下形式，m，n 分布为主题目标和微博文本的长度。

$$Target = \{w_1, w_2 ... w_m\} \tag{4-14}$$

$$Text = \{w'_1, w'_2 ... w'_n\} \tag{4-15}$$

（2）词向量嵌入层

词向量的嵌入层, 此层的功能是对输入的每一次词检索其词向量 (lookup 操作), 后续实验词向量的预训练由 GloVe 模型在大量无监督语料上训练可得, 预训练的词向量维度为 200, 且把词向量设置为可训练, 随神经网络模型的训练动态调整权重。经过此层后主题目标信息和微博文本信息可以转换成以下形式。

$$Target_v = \{y_1, y_2 ... y_m\} \tag{4-16}$$

$$Text_v = \{x_1, x_2 ... x_n\} \tag{4-17}$$

（3）主题目标编码与文本编码层

由于微博和 Twitter 文本立场分析中的主题目标包含的词语相对较少，为了减少模型的参数空间，因此编码主题目标信息的方法是平均所有词向量。微博或 Twiiter 文本包含了文本立场分析的绝大部分信息，双向的 GRU 单元能比单向更好抽取文本中的信息，对于当前词的隐藏状态包含前向和后向的 GRU 隐状态。具体的转换如公式所示。

$$q = \frac{\sum_m^{i=1} y_i}{m} \tag{4-18}$$

$$h_i^{\rightarrow} = GRU^{\rightarrow}(h_{i-1}^{\rightarrow}, x_i) \tag{4-19}$$

$$h_i^{\leftarrow} = GRU^{\rightarrow}(h_{i+1}^{\leftarrow}, x_i) \tag{4-20}$$

$$h_i = [h_{i-1}^{\rightarrow}, h_{i+1}^{\leftarrow}] \tag{4-21}$$

（4）注意力机制计算与基于注意力的隐藏状态合成

通过结合主题目标编码信息 q 和微博的文本信息的基于双向 GRU 模型的隐状态 h_i，分布计算主题目标对每个词的关注程度，对于文本分析影响大的词，注意力机制会给予大的权重，反过来对于无用信息，注意力机制则会给予小的权重。分别计算每个词的权重后，原隐状态点城注意力的权重形成基于注意力的隐藏状态，具体的计算公式如下

$$e_i = att(h_i, q) = W_n^t(tanh(W_{ah}h_i + W_{aq}q + b_a)) + b_m \tag{4-22}$$

$$a_i = \frac{exp(e_i)}{\sum_N^{n=1} exp(e_n)} \tag{4-23}$$

$$h_i' = a_i \odot h_i \tag{4-24}$$

（5）多窗口核卷积与池化操作

模型初始化不同大小的过滤器窗口, 每种尺寸的窗口随机初始化若干参数。图中所示窗口大小 w 为 2, 并随机初始化多个不同的过滤器。所得向量将送入修正线性单元 (Rectified Linear Units,ReLU) 等非线性目标激活函数中, 池化操作将卷积层中每个卷积过滤器所得长度为 N-h+1 的向量使用最大池化操作 (max pooling)，池化操作能减低参数维度，减少计算复杂度。具体公式如下

$$c_i = W * h[i : i + w - 1] + b \tag{4-25}$$

$$c = Max(c_1, c_2, ..., c_n) \tag{4-26}$$

（6）全连接分类层把每个卷积核经过池化操作的值拼接在一起作为模型最后的特征向量，对特征向量连接一个以三个输出单元的全连接层。全连接层输出外接 Softmax 归一化函数，输出的值为归一化到 0-1 之间的实数值，代表的意义为模型分类到某一个立场的概率，具体公式如下。

$$ouput_i = Softmax(W_{h3}C + b_{h3}) \tag{4-27}$$

R$R\ Max$

$$z_t = \alpha_g(W_z x_t + U_z h_{t-1} + b_z) \tag{4-28}$$

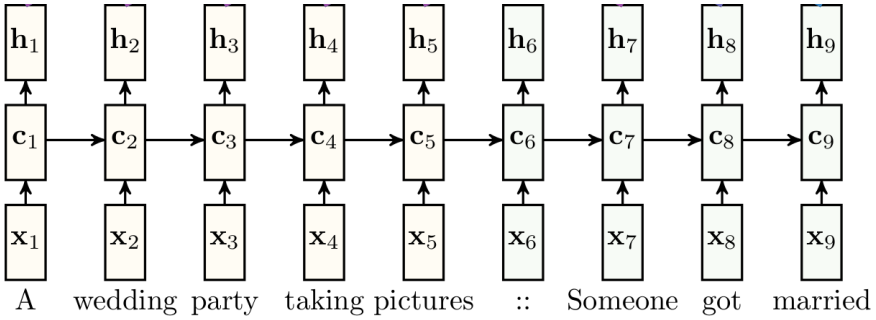

图 4-6 条件编码长短期记忆

参考文献

附录 A 带章节的附录

完整的附录内容，包含章节，公式，图表等

A.1 附录节的内容

这是附录的节的内容

附录中图的示例：

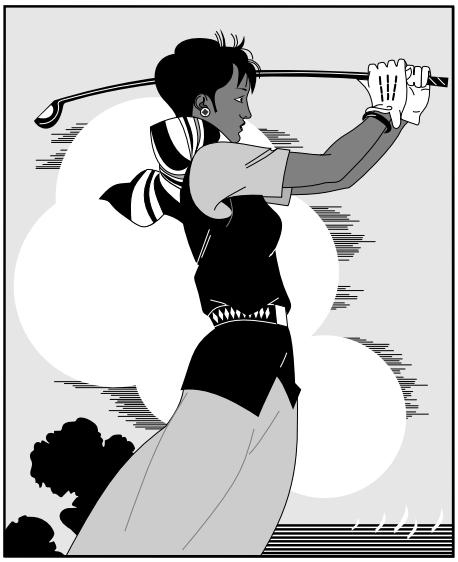

图 1-1 打高尔夫球的人

Fig. 1-1 The person playing golf

附录中公式的示例：

$$a = b \times c \tag{A-1}$$

$$E = mc^2 \tag{A-2}$$

附录 B 这个星球上最好的免费 Windows 软件列表

杀毒软件

avast! 免费杀毒软件——推荐
AVG 杀毒永久免费版——推荐
Avira Free Antivirus （小红伞）

攻读硕士学位期间发表的论文及其他成果

（一）发表的学术论文

[1] XXX，XXX. Static Oxidation Model of Al-Mg/C Dissipation Thermal Protection Materials[J]. Rare Metal Materials and Engineering, 2010, 39(Suppl. 1): 520-524.（SCI 收录，IDS 号为 669JS，IF=0.16）

[2] XXX，XXX. 精密超声振动切削单晶铜的计算机仿真研究 [J]. 系统仿真学报，2007，19（4）：738-741，753.（EI 收录号：20071310514841）

[3] XXX，XXX. 局部多孔质气体静压轴向轴承静态特性的数值求解 [J]. 摩擦学学报，2007（1）：68-72.（EI 收录号：20071510544816）

[4] XXX，XXX. 硬脆光学晶体材料超精密切削理论研究综述 [J]. 机械工程学报，2003，39（8）：15-22.（EI 收录号：2004088028875）

[5] XXX，XXX. 基于遗传算法的超精密切削加工表面粗糙度预测模型的参数辨识以及切削参数优化 [J]. 机械工程学报，2005，41（11）：158-162.（EI 收录号：2006039650087）

[6] XXX，XXX. Discrete Sliding Mode Cintrok with Fuzzy Adaptive Reaching Law on 6-PEES Parallel Robot[C]. Intelligent System Design and Applications, Jinan, 2006: 649-652.（EI 收录号：20073210746529）

（二）申请及已获得的专利（无专利时此项不必列出）

[1] XXX，XXX. 一种温热外敷药制备方案：中国，88105607.3[P]. 1989-07-26.

（三）参与的科研项目及获奖情况

[1] XXX，XXX. XX 气体静压轴承技术研究, XX 省自然科学基金项目. 课题编号：XXXX.

[2] XXX，XXX. XX 静载下预应力混凝土房屋结构设计统一理论. 黑江省科学技术二等奖, 2007.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《局部多孔质气体静压轴承关键技术的研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：　　日期：　　年　　月　　日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

致　谢

衷心感谢导师 XXX 教授对本人的精心指导。他的言传身教将使我终生受益。
感谢 XXX 教授，以及实验室全体老师和同窗们的热情帮助和支持！
本课题承蒙 XXXX 基金资助，特此致谢。
…