大数据情感分析综述

情感分析B组：梁仁杰，王宇，李志文，赵堃宇

**摘 要** 随着科技的进步和人类的物质文化需求的不断提高，计算机计算能力突飞猛进，人们对于高速、高效的获取数据中潜在的信息的需求日益提高。如今，大数据挖掘和大数据分析已经广泛应用于机器学习、深度学习、人工智能、信息挖掘等各个领域。其中，大数据在情感分析领域的作用近年来日益提高，本文首先简要介绍了文本情感分析的基本概念和应用方面。接着，本文着重从情感词典、机器学习和深度学习方面对情感分析的发展、方法、流程、优化等方面对国内外研究现状进行了详细的阐述和分析。（其中，情感词典方面分别从四种情感词典构建方法来进行阐述，机器学习方面{{{}}}，深度学习方面{{{}}}）。最后，对大数据情感分析的应用前景以及未来研究方向进行了展望。

关键词：大数据，情感分析，情感词典，机器学习，深度学习

A survey big data sentiment analysis

LIANG Ren-Jie，WANG Yu，LI Zhi-Wen，ZHAO Kun-Yu

**Abstract** With the advancement of science and technology and the continuous improvement of human material and cultural needs, computer computing capabilities have advanced by leaps and bounds, and people's demand for high-speed and efficient access to potential information in data is increasing. Today, big data mining and big data analytics have been widely used in machine learning, deep learning, artificial intelligence, information mining and other fields. Among them, the role of big data in the field of sentiment analysis has been increasing in recent years. This paper first briefly introduces the basic concepts and application aspects of text sentiment analysis. Then, this paper focuses on the development, methods, processes and optimization of sentiment analysis from the aspects of emotion dictionary, machine learning and deep learning, and elaborates and analyzes the research status at home and abroad. Finally, the application prospects of big data sentiment analysis and future research directions are prospected.

Keywords: big data, sentiment analysis, sentiment dictionary, machine learning, deep learning

1. 引言

随着科学技术的进步和发展，人们生活水平不断提高，人们对于生活高效、快捷的需求也日益加大，随之而来的便是大数据分析与信息挖掘。而目前，互联网行业的发展之迅速造就了互联网上数据量的庞大，大量的数据也就蕴藏着可用信息，情感分析就是从互联网上大量的文本数据中获取重要可用信息-情感的方法。其在微博情感获取，销售行业用户情感获取，舆论情感获取等方面都发挥着重要作用。

无论是使用构建情感词典还是使用深度学习、机器学习的方法来获取文本表露的情感，都需要对文本进行分词处理的工作，之后才能对文本进行分析获取文本情感。

在情感分析这一词刚刚出现时，情感分析领域普遍使用的方法是情感词典的方法，而且是人工构建情感词典的方法。随着技术的成熟，计算机计算能力的不断提升，出现了在人工构建的情感词典的基础上，采用自动化构建方法对已有的人工构建的情感词典进行扩展、优化，在机器学习、深度学习，并且对于情感分析精确度要求提升之后，又出现了使用机器学习、深度学习的方法对文本情感分类进行学习，进而通过学习结果进行情感分析。

本文第二节主要介绍了情感分析的概念、方法、分类、目前应用。第三节介绍了通过构建情感词典对文本进行请情感分析的方法。第四节介绍了基于机器学习的情感分析方法。第五节介绍了基于深度学习的情感分析析方法。最后对情感分析的未来研究方向进行了总结和展望。

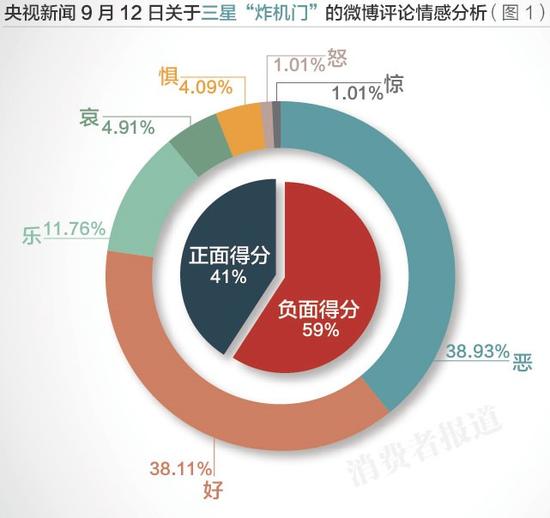
1. 情感分析的概念及应用

首先是对情感分析的一些相关概念，情感词典，情感分析等的介绍。

文本分词，

情感词典，是带有情感色彩的词或词组的集合，这些词可以使形容词，也可以是副词、名词或者是动词。情感词通常会带有某种情感极性，一般可分为正向情感词和负向情感词。正向情感词，一般表示带有积极、赞赏、肯定感情的词，也就是通常所说的褒义词。负向情感词一般表示带有消极、编译、否定感情的词，也就是通常所说的贬义词，如可怕、颓废、难过等等。

情感分析，又称为意见挖掘、倾向性分析，就是对带有情感色彩的主观性文本进行分析、处理、归纳和推理，最后得出文本的二元化或者多元化极性分类的方法。

其中，二元化情感极性分类，是将文本情感分为正极性和负极性，也就是表达出情感的好与不好。而多元化情感极性分类，是人为的对情感进行的分类，一般来说，有六大类，为乐、哀、惧、怒、惊、恶六类，但是在情感分析的时候，正向情感只有一类，相比起来，分类不算准确，所以多数情况下会在情感分类中增加一类，为好这一类，这样，正极性情感就有乐、好，负极性情感有哀、惧、怒、惊、恶，这么一来，情感的分类对于情感极性的把控就更为精准，也有利于将来的情感极性分析与分类的准确度。

对于文本的情感分类方法，传统意义上一般分为两类，一类是基于构建情感词典的方法，一类是基于学习的方法，而基于学习的方法里面又包括了，机器学习方法和深度学习方法。其中，基于构建情感词典的方法就是将情感词的情感极性进行权重计算、极性计算，得到的一个情感词的权重、极性列表，用构建出的情感词典对一篇已分好词的文本进行情感极性计算，从而获取文本的情感。基于学习的方法，包括基于机器学习的方法和基于深度学习的方法。

基于情感词典方法的一般流程为：首先获取大量的带有情感极性的文本，对文本进行分词，得到由词构成的文本，在词文本中搜索情感词，对情感词进行标记并进行计数，对搜索出的情感词进行权重标记，搜索情感词附近对于情感词情感强度有影响的程度副词，搜索情感词前的否定词，最后计算情感得分，从而构建出情感词典。然后将情感词典用于文本情感分析。

基于学习的方法的一般流程为：将已分好情感极性类别的文本，分为作为训练数据，输入到深度学习框架或者机器学习框架中，对文本进行特征提取，用提取出的特征进行模型训练，将训练好的模型用于文本情感分析。

无论是基于情感词典的方法还是基于学习的方法，其最后的目的都是能更搞笑更精准的对文本进行情感分类，无论是哪种方法，其作用和应用领域都大相径庭，有微博的情感分析、舆论情感分析、用户使用体验情感分析、主观评价情感分析等，其中，近些年出现了一些新的也更大的应用领域，比如说，对大众评价的情感分析来得出股票市场的涨跌情况，通过用户对商品的使用情况的反应的情感分析来制定新的营销策略，更甚有通过对人民群众的社交言论进行情感分析，来制定响应的政治策略。由此可见，文本情感分析的应用领域很多，很广，而且作用也很大，只是可能还没有被发现。

1. 基于情感词典的情感分析

基于情感词典的情感分析方法，最根本的就是情感词典的构建，而情感词典的构建步骤，分为如下几步，分词，情感词抽取，极性计算，权重计算，词典构建。其中情感词典抽取面临着巨大的挑战，首先，在社交网络中，用户往往会使用口语色彩浓重，隐晦模糊和非规范性表达的词语表达自己的观点，在这些非正式简单的文本中抽取情感词相当困难。另外，情感词极性会随着不同领域、不同语境发生变化，例如：在图书领域，“这本书里的故事太长了”，在电子领域，“这个电子产品的寿命真长”，他们的政府极性就不相同，这样的场景也增加了文本情感词提取的难度。由于这些原因，再加上要考虑文本语料的句子长短不一，内容规范性，领域性各个方面的问题等等这些原因的限制，现有的情感词典的构建方法会因为语料及任务的不同而采用不同的构建方法。

在处理网络文本的过程中，情感词的抽取及极性判断对于情感分析任务至关重要。大量的研究成果表明，情感词语的抽取和极性判别往往是一体化的工作，极性的分配会根据情感词抽取方法的不同而采取不同的策略。情感词典的构建分为三种：基于手工的方法、基于词典的方法和基于语料的方法，而本文将从技术角度对情感词典的构建进行综述。大体上，情感词典的构建从构建方式上进行分类可分为两类：人工构建的方法和自动构建的方法。从构建方法上进行分类大体上可分为四类：基于启发式规则的方法、基于图的方法、基于词对齐模型的方法和基于表示学习的方法。

* 1. 构建方式分类
     1. 人工构建情感词典

人工构建情感词典的方式，主要是利用大量现有的情感资源对前人总结的情感资源进行扩充标注，进而形成各种基础情感词典。其优点是便于创建更为丰富的词条信息，并且便于控制。

其中，大连理工大学的徐琳宏等通过手工情感分类和自动获取强度两种方法，构造了情感词汇本体。手工情感分类主要采取基于转换的错误驱动学习方法，自动获取词汇的情感强度用的是点互信息（Pointwise MutualInfor-mation，PMI）方法，计算公式如下：

 ⑴

其中，Ｗｕ表示具有ｕ类情感的词，Sui表示具有u类情感的第ｉ个标准词，计算 Ｗ 与所有具有ｕ类情感的标准词之间的互信息，选择互信息最大标准词的强度作为词汇 Ｗ 在ｕ 类情感上的强度。

王勇等人为了对中文微博进行情感分类，构建中文微博的极性词典。在各大微博网站随机抓取100000条微博，通过多次人工标注和校对，将2199个情感词进行正负向和强弱程度区分，根据微博表达的多样性，还构建了表情符号词典、否定词典和双重否定词典。

目前通过人工构建的情感词典主要有：哈佛大学的General Inquirer Lexicon、匹兹堡大学提供 的Opinion Finder主观情感词典、伊利诺伊大学BingLiu提供的词典资源、普林斯顿大学构建的英文情感词典 WordNet、台湾大学的中文情感极性词典（NTUSD）、知网情感词典HowNet等。

人工构建情感词典在扩充词条信息和操控便利性方面有一定优势，但是大大增加了人工开销，并且扩充范围有限，因此不适合跨领域研究。近年来自动构建情感词典的方法逐渐成为研究方向。

* + 1. 自动构建情感词典

通过自动构建情感词典，能很好地降低人工成本，并在一定程度上增强领域适用性，所以，近年来研究人员更多地致力于情感词典的自动构建工作。自动构建情感词典的方法主要有基于知识库的方法、基于语料库的方法以及基于知识库与语料库相结合的方法。

基于知识库的方法主要通过对现有知识库（如英文的WordNet、中文的HowNet）进行语义分析或扩展构成情感词典，以判断未知文本信息的情感倾向。如对WordNet进行扩展，加入名词、动词和副词，使情感词典更加全面。

基于语料库的方法主要是通过从大量语料中自动学习得到情感词典，并且通过对不同领域的语料进行提取，可以得到特定领域的情感词典。

如Hatzivassiloglou和McKeown提出一 种从大型语料库收集的间接信息自动检索语义取向信息的方法。该方法依赖于语料库，实现了高精度（超过90%），考虑了情感词或短语和特征词域的依赖关系，并在语料库更改时自动适应新域，可以直接应用于其它单词类。

Turney等使用 PMI方法扩展基本的正、负词汇，然后使用语义极性（ISA）算法分析情感文本，处理一般语料库数据的准确性率达到74%。考虑到用户行为，Yang等利用拉普拉斯平滑技术对SO-PMI算法进行改进，建立了中文酒店评论情感词典。其中PMI如式（1），然 后引入语义取向（SO），算法

如下：



⑵

将定义为正种子词集，为负种子词集，为正种子词数，为负种子词数。其中在计算时引入拉普拉斯参数ｔ。

周咏梅等人提出了一种中文微博情感词典构建方法。该方法利用上下文熵对微博中的网络用语进行判定，

公式如下：



⑶



⑷

其中，设为候选词的上文单字集合，为候选词的下文单字集合，和表示出现在候选词上文和下文的单字总频次，其中和分别为单字和出现在候选词 上、下文的频次。然后利用对网络用语进行过滤，其中是词在语料库中出现的次数，是语料库中的文档总数，是语料库中词的文档频次。

利用基于SO-PMI算法的情感强度计算方法筛选含有网络用语的微博，对这些微博进行人工标注，并计算情感极性和强度，构成词典。

Bravo-Marquez等人利用自动注释的推文构建Twitter意见词典，使用点互信息语义取向（PMI-SO）建模，并使用随机梯度下降语义取向（SGD-SO）学习词与情感之间的线性关系。

近年来由于互联网中的文本信息增长速度过快，一些网络词语出现，使得单纯利用原有知识库或互联网中的语料构建情感词典不能满足现有文本信息的要求，因此研究人员更倾向于利用知识库与语料库相结合的方法构建情感词典。通过将扩充的情感知识库及从语料库中提取的情感词汇引入情感词典，使构成的情感词典更加丰富。

杨小平等人利用Word2Vec工具，从海量的语料库中训练出一套词向量，并通过综合筛选 NTUSD词 典、知网情感词典和大连理工大学开发的情感本体库，构成 Sen-tiRuc词典，通过机器学习对情感色彩进行消歧，并对词典进行了同义关系优化、反义关系优化和句子级描述力优化，在通用领域数据集上取得了较好的实验结果。

基于词典的文本情感分析技术由于构建的词典往往只针对某个领域，对于跨领域情感分析的效果不够好，而且词典中的情感词可能不够丰富，对于短文本和特定领域文本进行情感分析的效果更好。因 此，对于长文本来说，更好的解决方法是利用机器学习方法。

* 1. 构建方法分类
     1. 基于启发式规则的方法

基于启发式规则的方法主要是通过观察大量语料的特性，找到一些语法模式、语法规则、语义特征和语言学特性，然后抽取出情感词并判断其极性。将基于启发式规则的方法分为两类：简单规则的方法和复杂规则的方法。

**简单规则方法：**

Hatzivassilogl和McKeown利用一些种子情感词和语言学特性来发现更多的情感词并判断其极性。他们的方法分析文档集中抽取出的由连词连接的形容词对，如连接 词And,but,either-or和neither-nor等。由and连接的形容词对往往具有相同的极性，如She is beautiful and clever，而由but连 接的形容词对往往具有相反的极性，如She is beautiful but selfish。他们利用语言学里的连词特性抽 取出了很多形容词对，但是无法抽取出语料中大量 单独的形容词。

Turney和 Littman定义了两个词之间的点间互信息(pointwise mutual information, PMI)，通过计算目标词与种子情感词之间的ＰＭＩ判断情感词及其极性。PMI的基本思想是两个词共现的次数越多，二者的关系就会越密切。两个词之间的PMI值，用式（5）计算。



⑴

其中表示两个词共同出现的概率，和表示两个词单独出现的概率。这种方法可以识别各种词性的情感词，缺点是需要辅助的网络资源。

有一群人（hu和liu等）首先找到频繁出现的产品特征（名词），然后在其附近找到相关的情感词，再利用WordNet中同义词和反义词关系判断候选词的极性。Kanayama and Nasukawa扩展了Hatzivassil- oglou的方法，提出了句子内部和句子之间的情感关联性思想。他们认为连续的若干句子往往具有相同的情感倾向，如果其中一个句子含有情感词，那紧连它的句子也会含有情感词并具有相同的情感极性。这种方法在上下文句子中没有种子情感词的情况下召回率会大大降低。

简单规则大致可以分为依据连词、依据名词、依据句子。

依据连词是利用一些种子情感词和语言学特性来发现更多的情感词并判断其极性。我们可以分析文档集中抽取出的由连词连接的形容词对，如连接词and、but、either、or和neither、nor，等。由and即“和”连接的形容词对往往具有相同的极性，如“上了大数据分析我觉得开心又充实”、“张华平老师博学又帅气”。当你知道了“开心”和“博学”的极性是好时，你就可以判断出“充实”和“帅气”的极性也是好的。而由but即“但是”连接的形容词对往往具有相反的极性，如“A虽然很帅但是他却是单身”，这里“帅”和“单身”就是一对相反的极性。这种方法利用语言学里的连词特性抽取出了很多形容词对，但是无法抽取出语料中大量单独的形容词。我们也可以通过计算目标词与种子情感词之间的定义好的的点间互信息（PMI）判断情感词及其极性。PMI的基本思想是两个词共现的次数越多，二者的关系就会越密切。这种方法具体如何计算上面已经有介绍，其可以识别各种词性的情感词，缺点是需要辅助的网络资源。

依据名词是指找到频繁出现的产品特征（名词），然后在其附近找到相关的情感词，再利用Wordnet中同义词和反义词关系判断候选词的极性。

依据句子是指依据句子内部和句子之间的情感关联性的思想。连续的若干句子往往具有相同的情感倾向，如果其中一个句子含有情感词，那紧连它的句子也会含有情感词并具有相同的情感极性。这种方法在上下文句子中没有种子情感词的情况下召回率会大大降低。就比如依据连词里面的两个句子可以在一段语境里面连起来，且他们的情感极性和四个情感词的极性都是相近的。

**复杂规则方法：**

复杂规则主要有双重传播、自动抽取和情感词典的扩展。

近年来，很多技术方法的逐渐兴起，人们将更多的规则应用到情感词典的构建中。许多观察发现，评价词（情感词）和评价对象（情感词所修饰的对象）往往是相互联系的，而他们的联系为情感词的抽取提供了很重要的信息。例如，“这个型号的相机非常漂亮”，如果我们知道“相机”是评价对象，那么修饰“相机”的形容词“漂亮”就会被认为是情感词。基于这种思想，我们可以利用评价词和评价对象的关系抽取情感词并判断其极性，这就是双重传播的思想，这种思想即联合抽取评价词和评价对象。我们可以借助依存句法、POS(Part of Speech)标注、parser结果来分析评价词和评价对象之间的关系，再根据定义的八条规则迭代扩展情感词集。这种方法大大增加了召回率，但在词典扩展的过程中由于引入了噪音导致准确率不够高，另外这种方法不适合处理网络上一些非正式的文本。也有人在总结研究学者的经典方法后构建了一个多步的方法，同时利用连词和双重传播的方法抽取情感词，并利用一些语言学模式进行词语的极性消歧。

由于手工构建规则的方法耗时耗力且比较有局限性，有研究学者采用自动抽取的方法。我们可以抽取unigram和bigram作为词典元素，利用PMI计算词典元素之间的关联度，构建了一个分布式情感词典，对于每个情感词都有一些与它相关联的情感词列表。这种情感词典在跨领域的情感分类任务上取得了较好的效果。

情感词典的扩展则是把情感词的词性从形容词和动词扩展到带有情感色彩的名词和名词词组上。这也是近年来用户在社交网络、帖子论坛、用户评论等网络平台发表带有主观情感色彩的文本的新趋向。例如，“舔狗”、“大佬”等名词都含有丰富的感情色彩和情感极性，我们在构建情感词典时也要把这些赋有情感的名词构建进去。

近年来，随着很多技术方法的逐渐兴起，人们将 更多的规则应用到情感词典的构建中。许多观察发 现，评价词（情感词）和评价对象（情感词所修饰的对象）往往是相互联系的，而他们的联系为情感词的抽取提供了很重要的信息。例如，“这个型号的相 机非常漂亮”，如果我们知道“相机”是评价对象，那么修饰“相机”的形容词“漂亮”就会被认为是情感词。基于这种思想，有人相处一种方法，沿袭了Kanayama和 Nasukawa的工作，他们利用评价词和评价对象的关系抽取情感词并判断其极性，提出了双重传播 （Doublepropagation）的思想，这 种 Bootstraping的思想联合抽取评价词和评价对象。他们借助依存句法、POS（PartofSpeech）标注、Parser结果来分析评价词和评价对象之间的关系，再根据定义的八条规则迭代扩展情感词集。这种方法大大增加了召回率，但在词典扩展的过程中由于引入了噪音导致准确率不够高，另外这种方法不适合处理网络上一些非正式的文本。Agathangelou等人在总结研究学者的经典方法后构建了一个多步的方法，同时利用连词和双重传播的方法抽取情感词，并利用一些 语言学模式进行词语的极性消歧。

由于手工构建规则的方法耗时耗力且比较有局限性，有研究学者采用自动抽取的方法。Bollegala等人抽取 Unigram和 Bigram作为词典元 素，利用PMI计算词典元素之间的关联度，构建了一个分布式情感词典，对于每个情感词都有一些与它相关联的情感词列表。他们构建的情感词典在跨领域的 情感分类任务上取得了较好的效果。Vishnu等利用多个领域的评论数据构建了特殊领域的情感词 典和领域无关的情感词典。还有些方法是把情感词的词性不仅局限于形容词和动词，还扩展到带有情感的名词和名词词组上。

基于启发式规则的方法优点是比较简单，针对性强，能够抽取特殊领域的情感词；缺点在于比较耗时，人工定义的规则也相对有局限性，可扩展性差， 在处理网络上那些非正式的文本时利用语法信息往往会产生很多错误。

**小结：**基于启发式规则的方法优点是比较简单，针对性强，能够抽取特殊领域的情感词；缺点在于比较耗时，人工定义的规则也相对有局限性，可扩展性差，在处理网络上那些非正式的文本时利用语法信息往往会产生很多错误。

* + 1. 基于图的方法

在情感词抽取的研究中，越来越多的研究学者倾向于使用基于图传播的方法抽取情感词并判断极性。算法过程通常分两步：

１）首先建立一个图，图的顶点是由待抽取的目标词或目标词组组成，边的权值为两个顶点之间的相似度；

２）然后在这个图上用一个图传播的算法迭代计算顶点的情感值。

已 有 的 图 传 播 算 法 有 随 机 游 走 （Random Walk）、PageRank、标 签 传 播 （Label Propagation）等。在计算边权值的过程中，会用到语法分析结果、句法上下文和词典中的语言学知识等信息。

**单一化方法：**

单一化情感图顶点主要分为最短路径、两步模型、情感词典网络图。

有学者研究语义指向利用WordNet构建了一个同义词网络，通过计算形容词与种子情感词good和bad的最短路径确定该形容词的语义指向。这种方法需要一个同义词词典，而词语之间的反义关系却不能包含在网络中，另外只能局限于形容词。之后又有学者改进了这种方法，他们利用 WordNet上词语的注释信息、同义词、反义词以及上位词信息构建了一个图，然后利用 Spin计算词语的极性。这种方法需要额外的词典资源。以上的方法都需要使用词典资源，然而，很多语种的情感资源并不像英文一样丰富。对于一些字典资源较稀缺的语种，有研究学者避开了词典资源的使用并取得了较好的抽取效果。

由于缺少俄语词典资源，一些学者为了从 Twitter流中抽取情感词提出了一个两步模型的方法，首先利用语言学和统计信息在电影领域训练了一个监督模型，然后在Twitter数据集上抽取特殊领域的情感词并利用SPin模型判断极性。他们的方法优点在于一旦训练就可以运用在不同领域和不同语言。

同样没有利用词典资源，有学者利用Web文档中的共现统计信息构建了一个情感词典网络图，然后使用一个图传播算法计算候选词的极性。这种方法优点在于未利用词典、ＰＯＳ标注以及语法分析相关的信息。

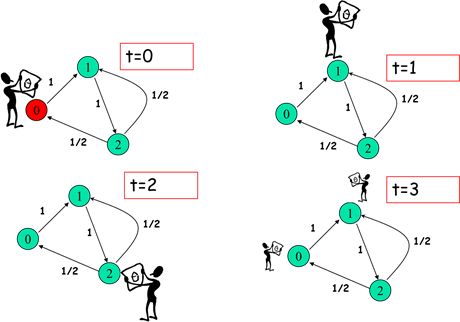
**多元化情感图顶点：**

很多研究学者发现情感词和评价对象的关系后，只把情感词和评价对象作为图模型的顶点，这样可能会导致准确度不高。鉴于此，一些研究学者在情感图顶点中加入一些辅助元素，例如，表情符号、 语义模式等。Ｘｕ等人提出了一个两级框架结构，首先他们构建了一个情感图游走算法，不仅将情 感词和评价对象作为图的顶点，还在情感图中加入语义模式，利用随机游走计算候选词的置信度，置信度越高，被认为是情感词的可能性越大；然后他们使用一个自学习策略过滤掉一些高频噪音并捕获一些长尾词。他们的方法能够提高准确率，过滤掉那些错误的情感关系和评价对象，但情感词的词性只局限于形容词。Feng等人把微博上的图形表情符号加入到情感图中，表情符号和候选的情感词共同作为图的顶点，然后随机游走算法抽取排序较前的情感词。这种方法不需要人工标注训练语料，也不用设计语法模式，但对于低频情感词的抽取具有局限性。

基于图的方法优点在于可以将词与词之间的各 种联系以特征的形式融入情感图中，另外利用图传 播的算法往往能抽取到大量的情感词，缺点在于在图传播算法过程中可能引进很多的噪音。如何优化图传播算法以及选择准确的词语间特征是未来的挑战。

图传播算法-随机游走：

随机游走（random walk）也称随机漫步，随机行走等是指基于过去的表现，无法预测将来的发展步骤和方向。核心概念是指任何无规则行走者所带的守恒量都各自对应着一个扩散运输定律，接近于布朗运动，是布朗运动理想的数学状态，现阶段主要应用于互联网链接分析及金融股票市场中。许多算法包括PageRank算法在内的很多链接分析算法都是建立在随机游走模型基础上的。



在最初阶段，用户打开浏览器浏览第1 个网页，假设我们有一个虚拟时钟用来计时，此时可以设定时间为1，用户在看完网页后，对网页内某个链接指向的页面感兴趣，于是点击该链接，进入第2 个页面，此时虚拟时钟再次计时，时钟走向字2，如果网页包含了k 个出链，则用户从当前页面跳转到任意一个链接所指向页面的概率是相等的。

用户不断重复以上过程，在相互有链接指向的页面之间跳转。如果对于某个页面所包含的所有链接，用户都没有兴趣继续浏览，则可能会在浏览器中输入另外一个网址，直接到达该网页，这个行为称为远程跳转（Teleporting）。假设互联网中共有m 个页面，则用户远程跳转到任意一个页面的概率也是相等的，即为1/m。随机游走模型就是一个对直接跳转和远程跳转两种用户浏览行为进行抽象的概念模型。

图传播算法-PageRank：

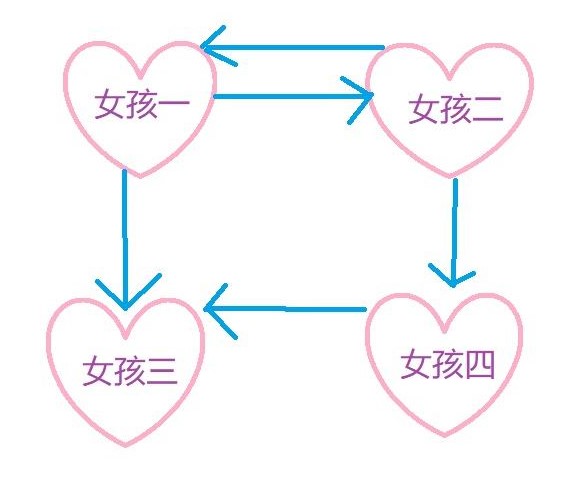
PageRank是Google专有的算法，来衡量特定网页相对于搜索引擎索引中的其他网页而言的重要程度。它由Larry Page和Sergey Brin在20世纪90年代后期发明。PageRank实现了将链接价值概念作为排名因素。

PageRank通过链接关系构建出图，每一个顶点设置同样的PageRank值，通过若干轮的计算，会得到每一个顶点所获得的终于PageRank值。随着每一轮的计算进行，顶点当前的PageRank值会不断得到更新。在一轮更新顶点PageRank得分的计算中，每一个顶点将其当前的PageRank值平均分配到本顶点包括的出链上，这样每一个链接即获得了对应的权值。而每一个顶点将全部指向本顶点的入链所传入的权值求和，就可以得到新的PageRank得分。当每一个顶点都获得了更新后的PageRank值，就完毕了一轮PageRank计算。 PageRank算法建立在随机冲浪者模型上，其基本思想是：网页的重要性排序是由网页间的链接关系所决定的，算法是依靠网页间的链接结构来评价每个页面的等级和重要性，一个网页的PR值不仅考虑指向它的链接网页数，还有指向’指向它的网页的其他网页本身的重要性。

PageRank具有两大特性：

1）PR值的传递性：网页A指向网页B时，A的PR值也部分传递给B。

2）重要性的传递性：一个重要网页比一个不重要网页传递的权重要多。

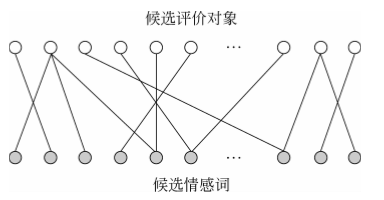


我们来用A同学举例说明一下方便大家理解。我们可以模拟一个场景：A同学由于单身太久想要脱单，我们都替他着急于是想要在相亲网上帮他找个对象。首先我们需要点开世纪佳缘或者珍爱网，点开女嘉宾信息，查看不同的女孩子，简单的根据对年龄、择偶要求、学历、颜值、这四个特征进行了初步筛选。我们在第一个女孩子所处网站停留10分钟，然后突然看到还有更适合的女孩子，所以从这个网站跳转到第二个女孩子的网页上（第一个网站加10分），在第二个女孩子处看到择偶要求上写着不要程序员，这个就不能忍了，不要我们我们也不理她，于是在这个停留了5分钟的网站跳转到了下一个女孩子网站（这个网站加5分）。我们一看这个女孩子挺好的，收藏了就她了，于是就没有再去其他网站了。（这里为特殊情况，即这个女孩子不指向其他女性，则不满足收敛性），但是我们这帮直男看女孩都太简单了，况且还是给A脱单，还是要他自己看看，增加个看不看得上的概率（阻尼系数：0.8），即可计算每个女孩的得分了。

**小结：**基于图的方法优点在于可以将词与词之间的各种联系以特征的形式融入情感图中，另外利用图传播的算法往往能抽取到大量的情感词，缺点在于在图传播算法过程中可能引进很多的噪音。如何优化图传播算法以及选择准确的词 语间特征是未来的挑战。

* + 1. 基于词对齐模型方法

研究人员发现基于语法的方法在处理非正式的文本时往往会产生很多错误，另外这种方法在处理小或中等大小的语料时能取得较好的结果，但在处理大语料数据时往往会遇到瓶颈。鉴于此，有学者提出了利用机器翻译中的词对齐模型挖掘情感词和评价对象之间的关系。



如图所示，黑色顶点表示候选情感词，白色顶点表示候选评价对象，具体步骤如下：

１）构建二分图的顶点，选取所有的形容词作为候选情感词顶点，所有的名词或名词词组作为候选评价对象顶点；

２）利用机器翻译模型IBM挖掘情感词和评价对象之间的关系；

３）采用基于图的算法计算候选词的置信度，置信度越高，被选择的概率越大。

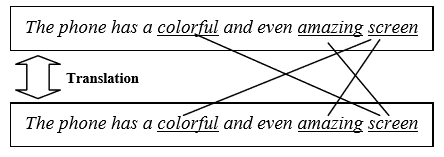
**挖掘关联：**

在意见挖掘中，一个基本问题是意见目标提取。此任务是提取表达意见的项目。在评论中，意见目标通常是名词/名词短语。例如，在“电话具有丰富多彩甚至惊人的屏幕”的句子中，“屏幕”是意见目标。在线产品评论中，意见目标通常是产品或产品特征，因此该任务在以前的工作中也被称为产品特征提取。

为了提取意见目标，许多研究认为意见词是强有力的指标，这是基于观察词通常位于意见目标周围的观察，并且它们之间存在关联。因此，大多数过去的方法根据意见词和意见目标之间的关联迭代地提取意见目标。例如，“彩色”和“惊人”通常用于修改关于手机的评论中的“屏幕”，因此它们之间存在强烈的关联。如果已知“多彩”和“惊人”是意见词，“屏幕”很可能是该领域的意见目标。此外，提取的意见目标可用于根据其关联扩展更多意见词。这是一个相互强化的过程。

因此，意见目标和意见词之间的挖掘关联是意见目标提取的关键。为此，大多数先前的方法，被命名为相邻方法，采用相邻规则，其中意见目标被认为具有意见关系与给定窗口中的周围意见词。然而，由于窗口大小的限制，不能准确地捕捉意见关系，特别是对于长距离关系，这会损害估计意见目标和意见词之间的关联。为了解决这个问题，一些研究利用了依赖树等句法信息。如果意见词和意见目标之间的句法关系满足设计模式，那么它们之间就存在意见关系。实验一致地报道，基于语法的方法可以产生比小型或中型语料库的相邻方法更好的性能。基于语法的方法的性能在很大程度上取决于解析性能。但是，在线评论通常是非正式的文本（包括语法错误，拼写错误，不正确的标点符号等）。因此，解析可能会产生许多错误。因此，对于包括大量非正式文本的来自Web的大型语料库，这些基于语法的方法可能遭受解析错误并引入许多噪声。此外，这个问题在非英语语言评论中可能更严重，例如中文评论，因为在这些语言上解析的表现往往比英语差。

为克服上述两种方法的不足，有学者提出了一种新的无监督方法，通过使用基于单词的翻译模型（WTM）来提取意见目标。首先制定意见目标和意见词之间的意见关系，作为一个词对齐任务。这种方法认为意见目标可以通过单语词对齐找到相应的修饰语。

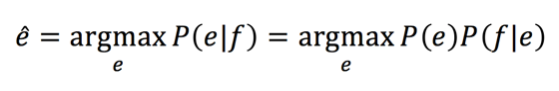


如图所示，意见词“多彩”和“惊人”通过词对齐与目标“屏幕”对齐。为此，我们使用WTM对意见目标和意见词之间的挖掘关联执行单语词对齐。在这个过程中，可以全局考虑几个因素，例如单词共现频率，字位置等。与相邻方法相比，WTM不识别给定窗口中单词之间的意见关系，因此可以有效地捕获长跨关系。与基于语法的方法相比，WTM可以在不使用解析的情况下有效避免解析非正式文本时的错误。所以它会更强大。此外，通过使用WTM，我们的方法可以捕获“一对多”或“多对一”关系（“一对多”意味着，在一个句子中，一个意见词修改了几个意见目标，和“多对一”意味着几个意见词修改一个意见目标）。

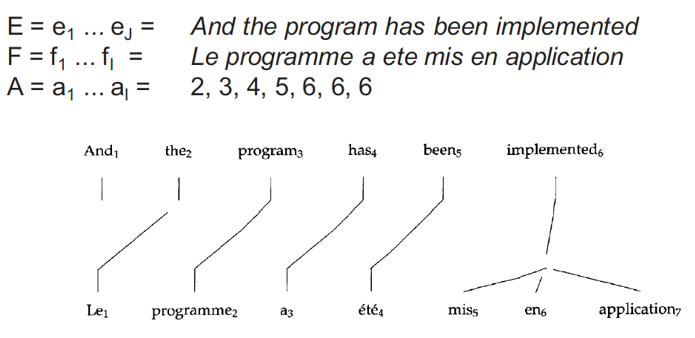
**IBM模型介绍：**

基于单词的翻译模型的基础是IBM模型，IBM model是统计机器翻译中的经典翻译模型，也是基于词的统计机器翻译系统的基础。IBM翻译模型共有5个复杂度依次递增的统计翻译模型，IBM model1是其中最简单的模型，也是其他模型进行计算的基础。IBM模型1仅仅考虑词对词的互译概率，模型2引入了词的位置变化概率，而模型3加入一个词翻译成多个词的概率。

假设任意一个英语句子e和法语句子f，定义f翻译成e的概率为P(f | e) ，于是将f翻译成e的问题则变成求解如图4.3方程的问题：



作为一个机器学习任务我们首先要有一些训练数据。而对于机器翻译工作而言，所谓的训练数据就是一些对应的双语互译语料，我们基于这些训练数据来估计 P(f | e)。而在这个过程中必须考虑两个问题：对齐和内容。



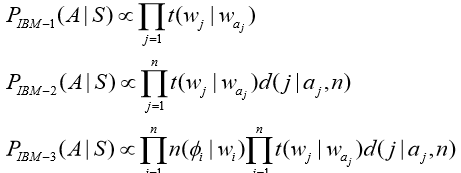
如图所示，对齐即哪个FJ是EI的翻译， 内容则是哪个词被选为FJ。

我们使用基于单词的翻译模型来执行单语单词对齐，这种对齐已被广泛用于许多任务，例如搭配提取、问题检索等等。如果我们直接将这种对齐模型用于我们的任务，名词/名词短语可以与除形容词之外的无关词语对齐，如介词或连词等。 因此，在对齐过程中，我们引入了一些约束：

1）名词/名词短语（形容词）必须与形容词（名词/名词短语）或空词对齐;

2）其他词只能与自己对齐。

总的来说，我们使用图4.5中3个WTM（IBM 1~3）来确定意见关系。



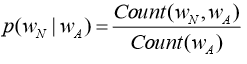
1）t模拟语料库中两个单词的共现信息。 如果形容词在评论中经常与名词/名词短语共同出现，则该形容词与该名词/名词短语具有高度关联。例如，在手机评论中，“大”通常与“手机大小”共存，因此“大”与“手机大小”高度相关。

2）d拟单词位置信息，其描述位置aj的单词与位置j中的单词对齐的概率。

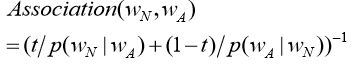
3）n模拟单词的生育能力，它描述了单词“一对多”对齐的能力。 ∮i表示与Wi对齐的单词数。 例如，“Iphone4具有惊人的屏幕和软件”。 在这句话中，“惊人”用于修饰两个词：“屏幕”和“软件”。 所以∮等于2为“惊人”。

**候选置信度估计:**

当我们已经提取出了名词-形容词对，就要开始找到他们之间的关联。从评论语句中收集所有单词对后，我们可以估计名词/名词短语和形容词之间的翻译概率，即



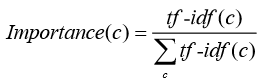
名词/名词短语和形容词之间的关联估计可以用下面公式来计算



然后我们计算每个意见目标候选人的置信度并对其进行排名。信心较高的候选人被视为意见目标。我们认为候选人的信心取决于两个因素：意见相关性和候选人重要性。

意见相关性反映了候选人与意见词相关联的程度。 如果一个形容词具有更高的置信度作为一个观点词，它所修饰的名词/名词短语将更有信心成为一个意见目标。类似地，如果名词/名词短语具有更高的置信度作为意见目标，则修改它的形容词很可能是意见词。这是一个迭代强化过程，表明现有的基于图的算法是适用的。

候选人重要性反映了候选人在语料库中的显着性。 我们根据其tf-idf即图4.8中计算公式将重要性分数分配给意见目标候选者f，其通过所有候选者的tf-idf分数的总和进一步归一化。其中c代表候选者，tf是数据集中的术语频率，df是使用Google n-gram语料库计算的。



基于这些关联，我们利用基于图的算法来计算每个意见目标候选者的置信度。









如图所示，我们使用C来表示候选置信度向量，即n \* 1向量。我们用候选者重要性分数设置候选者初始置信度，即C^0= S，其中S是候选初始置信度向量。名词/名词短语和形容词之间的边缘表示它们之间存在意见关系。边缘上的权重表示它们之间的关联，这是通过使用WTM估计的。M是意见相关矩阵，m\*n矩阵，其中，Mij是名词/名词短语i和形容词j之间的相关权重。 为了考虑候选人的重要性分数，我们引入了重新分配条件：在每个步骤中将候选意见相关性与候选人重要性相结合。 因此，我们可以得到候选置信度的最终递归形式如下。其中λ =[0,1] 是候选人置信度中候选人重要性的比例。 当λ=1时，候选置信度完全取决于候选者的重要性; 当λ=0时，候选置信度由候选意见相关性决定。S中的每个项目是使用Importance公式来计算的。

然后将具有较高置信度分数的候选者提取为意见目标。

**小结：**

基于词对齐模型的方法在情感词典构建上比较新颖，它能够有效地避免在网络非正式文本上的语法分析错误，另外 比较适合 处 理中等大 小 的语料。如何优化词对齐模型以及增加情感词的词性（如名词、动词等）到二分图中是值得深入研究的问题。

* + 1. 基于表示学习方法

近年来，词或词组的分布式向量表示在推动自然语言处理中取得了显著的进步，它能够获取大量准确的语法和语义关系。有学者提出了利用Skip-gram预测词或词组的上下文词并学习到词或词组的向量表示，即词语向量，如“漂亮”一词的词语向量是［0.782，-0.177，-0.106，0.109，-0.542…］，对于给定的词或词组，利用它的词语向量ei预测其上下文词。基于表示学习的方法具体步骤如下：

1）扩展Skip-gram模型.

2）开发具有混合损失函数的专用神经结构。

3）结合来自文本的情感极性的监督。

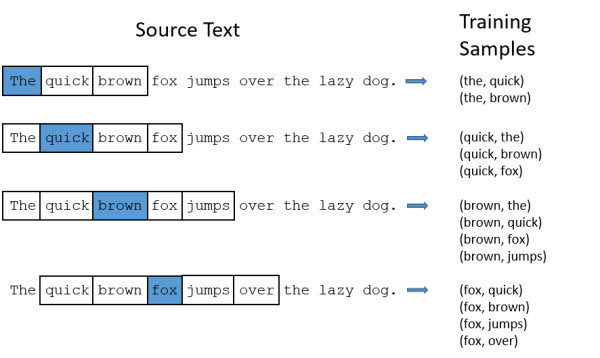
4）用分类器预测词汇表中每个短语的情感分数。

**Skip-gram:**

kip-gram即先基于训练数据构建一个的神经网络模型，但是当这个模型训练好以后，我们并不会用这个训练好的模型处理新的任务，我们真正需要的是这个模型通过训练数据所学得的参数，例如隐层的权重矩阵——后面我们将会看到这些权重在Word2Vec中实际上就是我们试图去学习的“word vectors”。就比如古典故事“买椟还珠”，说的就是这个意思，这里的最终输出结果就是“珠”，训练得到的隐层权重矩阵就是“椟”，我们要的是“椟”而不是“珠”。

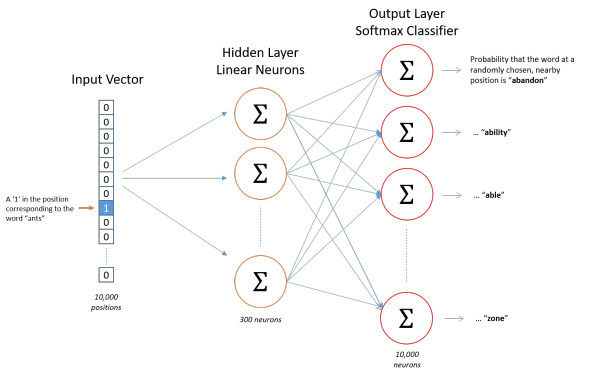


我们提到，训练模型的真正目的是获得模型基于训练数据学得的隐层权重。为了得到这些权重，我们首先要构建一个完整的神经网络作为我们的“Fake Task”，后面再返回来看通过“Fake Task”我们如何间接地得到这些词向量。接下来我们来看看如何训练神经网络。比如“The quick brown fox jumps over lazy dog”这个句子。我们可以从下图5.2中看到这个句子的输出结果，蓝色代表input word，设定skip\_window为2，该参数代表着我们从当前input word的一侧（左边或右边）选取词的数量，num\_skips也为2，该参数代表着我们从整个窗口中选取多少个不同的词作为我们的output word。



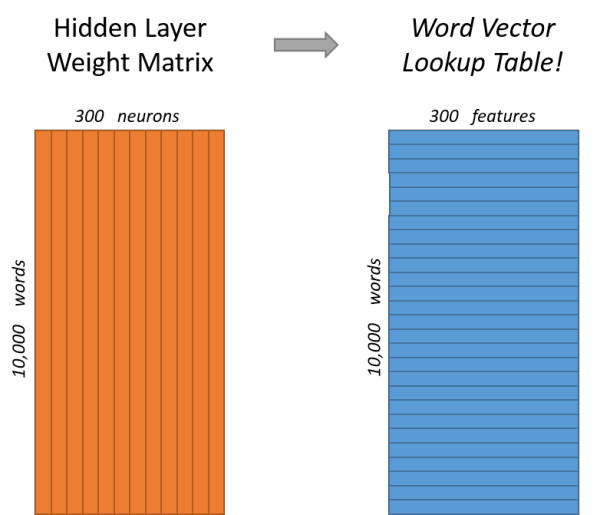
那么我们如何来表示这些单词呢，我们都知道神经网络只能接受数值输入，我们不可能把一个单词字符串作为输入，因此我们得想个办法来表示这些单词。

最常用的办法就是基于训练文档来构建我们自己的词汇表（vocabulary）再对单词进行one-hot编码。假设从我们的训练文档中抽取出10000个唯一不重复的单词组成词汇表。我们对这10000个单词进行one-hot编码，得到的每个单词都是一个10000维的向量，向量每个维度的值只有0或者1，假如单词ants在词汇表中的出现位置为第3个，那么ants的向量就是一个第三维度取值为1，其他维都为0的10000维的向量。



如图所示，这就是skip-gram神经网络的结构，模型的输入如果为一个10000维的向量，那么输出也是一个10000维度（词汇表的大小）的向量，它包含了10000个概率，每一个概率代表着当前词是输入样本中output word的概率大小。

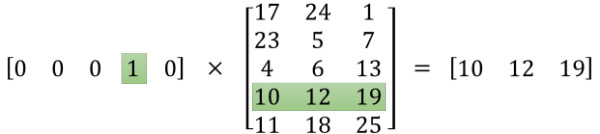
说完词汇的编码和训练样本的选取，我们来看下我们的隐层。



如果我们现在想用300个特征来表示一个单词（即每个词可以被表示为300维的向量）。那么隐层的权重矩阵应该为10000行，300列（隐层有300个结点）。

如图5.4中，左右两边分别从不同角度代表了输入层-隐层的权重矩阵。左图中每一列代表一个10000维的词向量和隐层单个神经元连接的权重向量。从右边的图来看，每一行实际上代表了每个单词的词向量。

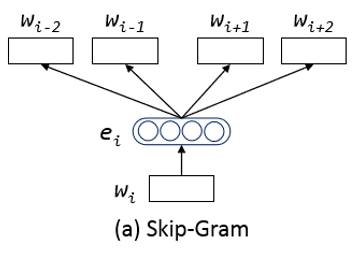
上面我们提到，input word和output word都会被我们进行one-hot编码。仔细想一下，我们的输入被one-hot编码以后大多数维度上都是0（实际上仅有一个位置为1），所以这个向量相当稀疏，那么会造成什么结果呢。如果我们将一个1 x 10000的向量和10000 x 300的矩阵相乘，它会消耗相当大的计算资源，为了高效计算，它仅仅会选择矩阵中对应的向量中维度值为1的索引行（这句话很绕），如图所示。

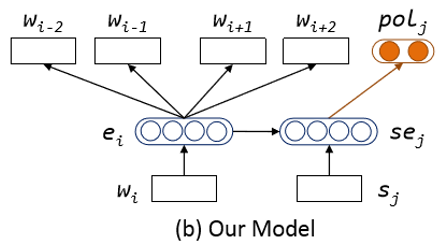


可以看到矩阵的计算的结果实际上是矩阵对应的向量中值为1的索引，上面的例子中，左边向量中取值为1的对应维度为3（下标从0开始），那么计算结果就是矩阵的第3行（下标从0开始）—— [10, 12, 19]，这样模型中的隐层权重矩阵便成了一个”查找表“，进行矩阵计算时，直接去查输入向量中取值为1的维度下对应的那些权重值。隐层的输出就是每个输入单词的“嵌入词向量”。

**扩展的基于表示学习的Skip-Gram模型:**

在传统的Skip-Gram模型的基础上，基于表示学习的方法扩大一小部分情感种子到收集培训数据以构建短语级别的分类器。具体而言，通过将情感信息和句法语境编码成短语的连续表示来学习情感特定短语嵌入（SSPE），这是一种低维，密集和实值向量。结果，SSPE的嵌入空间中的最近邻居有利于具有相似的语义使用以及相同的情感极性。然后扩展现有的短语嵌入学习算法，并开发具有混合损失函数的专用神经结构，以结合来自文本的情感极性（例如推文）的监督。从推文中学习SSPE，利用包含正面和负面表情符号的大量推文作为训练集而无需任何手动注释。为了获得更多用于构建短语级别情感分类器的训练数据，利用Urban Dictionary 2中的类似单词来扩展一小部分情绪种子，该单词是一个众包资源。最后，则利用分类器来预测SSPE词汇表中每个短语的情感分数，从而产生情感词典。





**小结：**

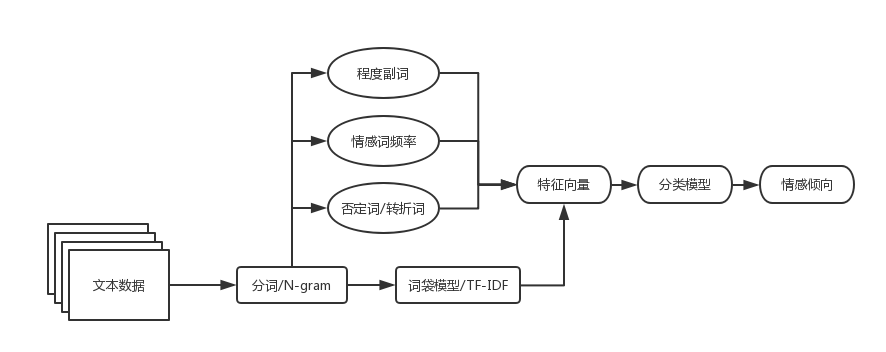
基于表示学习的方法是在近年来逐渐兴起的词语向量研究上发展起来的，具有广阔的研究前景，未来如何优化神经网络模型以及加入有效的情感特征是值得研究的方向。

1. 基于机器学习的方法

传统的情感词典中的情感词有较大程度的领域依赖性、时间依赖性和语言依赖性，同一词汇在不同的领域、时间和语言环境中可能会表达完全不同的情绪。另外可能还存在词典中情感词固定，难以及时捕捉新词、变形词等的缺陷。虽然基于规则的情绪方法可以在较短时间内获得分类结果，且可以加入一些规则来提高情绪分类的准确率，但在数据量较大时，规则的维护比较复杂，而且不易扩展。在大数据背景下，机器学习方法在分类，回归预测等领域具有很好的应用前景，情感分析作为一种特殊的分类应用，使用机器学习方法可以取得很好的效果。基于机器学习的情感分析方法主要分为有监督方法和半监督方法两类。

* 1. 基于机器学习的情感分类流程

使用机器学习进行情感分析的流程如下图所示，首先会输入具有情感分类的样本集，然后通过词袋模型将其分词生成词向量，基于机器学习的方法中，特征的提取和选择对预测的结果具有至关重要的作用，往往会选择分词之后的程度副词、情感词频率、否定词、转折词及TF-IDF值计算等。特征提取之后组合形成特征向量，然后选择传统的朴素贝叶斯算法、SVM算法等机器学习算法训练数据生成用于情感分类的模型。之后将待分类的文本输入到分类模型，预测其情感倾向。



* 1. 情感分析的特征提取

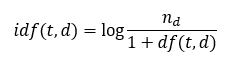
特征选取是否合适是影响有监督学习分类效果的一个主要因素，现有方法中用于情感分类的特征主要有词级、句子级和篇章级，其中词级包括词频（如词袋特征和词频-逆文档频率特征）、词性（如名词、动词、连接词）、语义（如词向量的相似度）、表情符合及其组合。

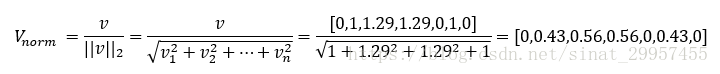
* + 1. 词袋模型

词袋模型即Bag-of-words model最早出现在自然语言处理（Natural Language Processing）和信息检索（Information Retrieval）领域。该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。词袋模型使用一组无序的单词(words)来表达一段文字或一个文档。该模型计算词向量的过程是，首先为文档中所有出现的词进行排序并赋予唯一索引，然后统计每次词在文档中出现的频率，并将其表示为向量。

使用词袋模型的方法来构建词向量的时候可能会遇到一个问题，有些单词在不同类型的文档中都出现的时候，这种类型的单词其实并不具备文档类型的区分能力，此时引入TF-IDF算法，TF-IDF算法即词频-逆文档频率算法，其计算方法如下，其中tf即上面词袋算法得到的词语在文档中出现的频率，idf值即逆文档频率，其求解方法中n\_d表示文档总数，df表示包含该单词的文档个数，然后再进行归一化。该算法的思想是当该词在文本中出现的频率较高，且在其他文本中出现较少时，该词对文本的分类就有很重要的作用。

https://img-blog.csdn.net/201804122233507?watermark/2/text/aHR0cHM6Ly9ibG9nLmNzZG4ubmV0L3NpbmF0XzI5OTU3NDU1/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70

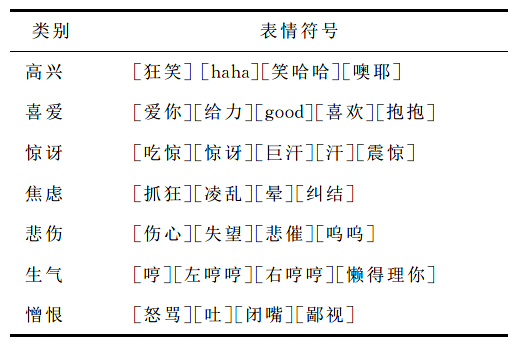




* + 1. 情感特征提取

现有方法中在词级别的特征提取包括依照情感词典的情感词频率特征、否定词、转折词及程度副词等在文本中具有明显情感倾向的词语，在早期的研究中，就有通过选取故事首句、特定连接词等30种特征将22篇格林童话划分为happy, sad, fearful, angry, disgusted, surprised, non-emotion等7类情感，最终通过SVM与SNoW结合的方法，达到准确率69.37%。将上面的情感词与词袋模型、TF-IDF算法计算的词向量相结合可以取得比较好的分类效果。对于微博等社交媒体短文本进行情感倾向的分析一直是情感分析种比较重要的研究内容，在社交媒体中，除了上面描述的常用情感词特征之外，表情符号一直是非常重要的组成部分，而且考虑到微博内容短小，字数限制往往在140字以内，所以微博中出现的表情符号往往能够代表整条微博的情感标签，所以在对微博等社交媒体的情感分析中，表情符号是很重要的情感特征。

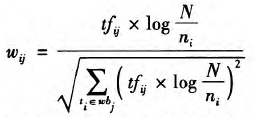
基于表情符号的情感分析首先会将常用的表情符号映射到需要分类的情感类别当中，如下表，基于这些已经标注了情感倾向的表情符号就可以借鉴之前传统的情感词典的方法，统计每个微博当中属于不同类别情感的表情符号个数，最终给一个关于整条微博的情感倾向标签。当然也可以将其中各类表情符号的个数作为提取的特征，作为后续机器学习方法的分类依据。



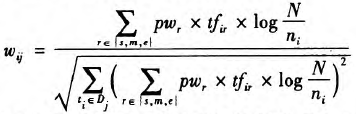
Zhao等人建立的基于表情符号与词袋相结合的情感分类系统Moodlens，将1000多个表情符号选取其中最常用的95个，将其映射到4个情感分类当中即愤怒、厌恶、开心和悲伤。其实验将收集到的韩这些表情符号的350万微博作为测试集，使用传统的朴素贝叶斯分类器，可以实现在线实时的情绪监控，准确率可达64.3%。

短文本分类是文本分类中的一个分支，短文本的情感分析也是情感分析当中很重要的一部分。由于文本长度短，特征稀疏等问题，导致传统的分类方法对其分类效果不好，针对这种短文本的结构短小、情感表达一致等特点，有学者提出通过CRF（条件随机场）模型提取短文本的情感特征，加入到分类器中，去把握整体句子的情感倾向。由于短句的情感具有一致性特点，即短句的某一个特征倾向于某一个情感，其上下文特征也倾向于该情感，这与CRF模型存在特征之间的约束特性高度一致，结合了最大熵模型和隐马尔可夫模型的特点，是一种无向图模型，近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。条件随机场是一个典型的判别式模型。该方法首先根据中英文标点符号将文本划分为短句，采用CRF模型在短句情感分类上进行建模，随后提取出长句中每一个短句的情感特征，加入到特征集合中去，再采用支持向量机等模型分类器对短句之间的情感进行建模，从而挖掘短句与短句之间的语义情感关联关系来达到较高的情感分析准确率。基于以上的方法，使用CRF对短句的分类结果准确率可达到78.35%，将短句特征加入到长句分类中通过SVM分类器准确率可达到85.39%。

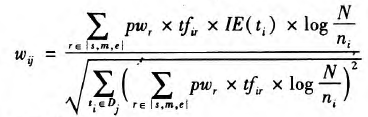
特征权重是用来衡量某个特征项在文本表示中的重要程度或区分能力的强弱。目前最常用的TF-IDF算法归一化后的公式如下：



其通过词频与逆文档频率来计算文档中每个词在文本中的重要程度。研究中发现文本中的首句、中间部分和尾句的极性对整个文本的极性贡献是不相同的，尤其对于微博等短文本来说，文本的首句、尾句往往对整个文本的极性贡献较大，而在标准的TF-IDF算法中忽略了这种位置不同带来的差异。因此有学者提出对上面的TF-IDF算法引入位置系数，将微博文本(〖wb〗\_j)分为首句(p\_s)、中间部分(p\_m)以及尾句(p\_e)三部分，在计算特征权重时，每部分赋予一定权值〖pw〗\_r，用于表示该部分特征对整个文本的贡献程度。将位置系数引入TF-IDF算法之后的计算公式变成：



特征项本身是有极性的，这样的极性往往对判定整个文本的情感倾向有很重要的作用，所以将特征项的极性值作为权重计算的一部分，进一步改进TF-IDF算法，加入位置权重系数和特征极性值的TF-IDF算法变成：



词级特征可以将大多数信息表示成词向量形式，并且可以较方便地衡量两个词之间的相似度，在情感分析工作中至关重要，然而事实上句子中的词语并不是相互独立的，不同的句法也会有不同的情感表达，所以词级的特征缺乏对文本预料整体上的考虑，在复杂句式中的分类可能就会存在准确率不高的问题。将上面的短句情感极性特征加入到分类特征中去可以很好地提高情感分类的准确率。

* + 1. 特征选择方法

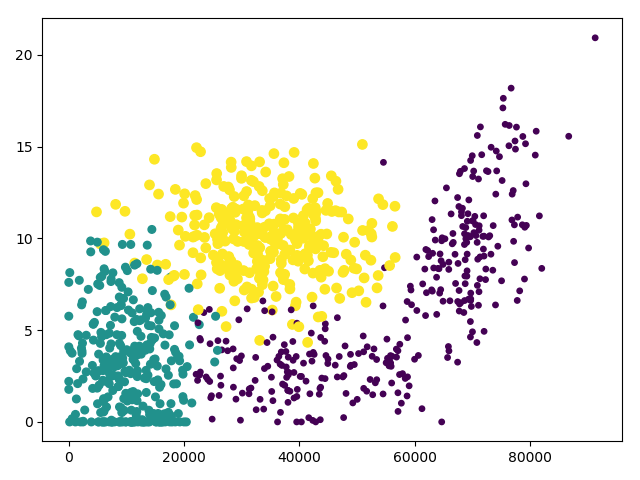
特征选择对于机器学习过程具有非常重要的作用，选取合适的特征选择方法有助于提升模型的性能，并帮助我们理解数据的结构。特征选择方法的主要功能是减少特征数量、降维，使模型泛化能力更强，减少过拟合，增强对特征和特征值之间的理解。常用的特征选择方法有单变量特征选择如卡方检验、Pearson相关系数、距离相关系数等、去掉取值变化小的特征等方法。训练分类器之前可以通过IG（信息增益）与CHI（卡方检验）特征选择方法对提取的特征进行降维和选择。

* 1. 机器学习方法分类

基于机器学习的情绪分类方法主要通过文本数据获取有价值的情感信息，按照方法的不同可以分为有监督的方法和半监督的方法。其中有监督的方法在对数据进行特征提取之后获得特征向量，之后就将其放到特定的机器学习算法中训练出分类器，然后就可对特定文本进行情感分析。根据分类算法的不同，可以将文本情感分析分了基于朴素贝叶斯（Naïve Bayes，NB）的方法、基于最大熵（Maximum Entropy，ME）的方法和支持向量机（SVM）的方法，半监督的方法则需要通过少量已标注数据来实现对大量数据的自动标注并最终训练一个情感分类模型。

* + 1. 基于朴素贝叶斯的方法

朴素贝叶斯是一个概率模型，其广泛应用于多个领域的工作中，在朴素贝叶斯技术中基本的想法是通过单词和类别的联合概率找出给定文本文档属于特定类别的概率，其在文本情感分类中有着很重要的应用。朴素贝叶斯算法就是求解新的数据属于每个类别的概率问题。



朴素贝叶斯算法是贝叶斯分类算法的一种，该算法以贝叶斯定理为基础，实现文本情感分析的分类任务。所谓贝叶斯定理就是在已知先验概率和A,B概率的前提下求解后验概率的过程，贝叶斯公式的具体形式如下：



朴素贝叶斯算法的算法思想是：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。于是其主要任务变成求解新的数据属于任一分类的概率P(y\_i |x)的问题，而通过贝叶斯定理可将其转换为求解P(x|y\_i)、P(y\_i)及P(x)的问题。此时朴素贝叶斯算法会假设组成数据的每个特征之间是相互条件独立的，于是就可以通过训练集求解所需的先验概率：



为防止求解先验概率时为零的情况，可以对其进行拉普拉斯平滑。

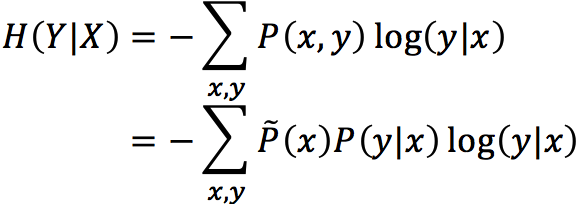
基于朴素贝叶斯算法的文本情感分析可以应用于许多领域，Soelistio等提出了一种使用朴素贝叶斯分类法分析数字报纸情感极性的简单模型，将其应用在数字报纸上进行政治情绪分析，从数字新闻文章中获取有关特定政治家积极或消极的情绪信息。Wikarsa等研究了一种使用朴素贝叶斯方法对Twitter用户进行情感分类的应用。Dey等利用朴素贝叶斯算法和KNN算法对电影评论和酒店评论进行情感分析，发现在电影评论中朴素贝叶斯比KNN效果好，但在酒店评论中，两者准确度差别不大。

基于朴素贝叶斯的文本情感分析技术是通过计算概率对文本情感进行分类，适合增量式训练，而且算法比较简单，对小规模数据表现良好。但该方法对输入数据的表达形式很敏感，而且需要计算先验概率，因此会在分类决策方面存在错误率。

* + 1. 基于最大熵模型的方法

最大熵模型属于指数模型类的概率分类器，基于最大熵原理，从适合训练数据的所有模型中选择具有最大熵的模型。所谓最大熵原理即假设对一个随机事件的概率分布进行预测时，预测应当满足全部已知的约束，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小，因此得到的概率分布的熵是最大。

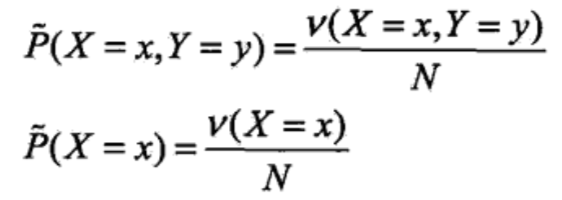
最大熵模型用数学方式表示即利用最大熵原理选择一个好的分类模型，使得条件概率的熵最大，即条件熵H(Y|X)：



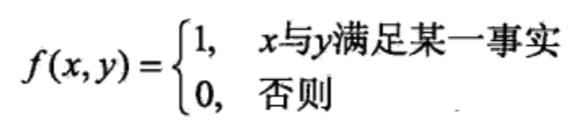
在没有任何约束条件的情况下，当且仅当输出的分类结果属于各个类别的概率相同时条件熵最大，但是在正常情况下要在满足已知条件约束的基础上实现熵最大。对于给定的训练数据集可表示如下：

https://img-blog.csdn.net/20180106134401457

对于给定的数据集来说，可以计算其联合概率分布P(X,Y)的经验分布P^~ (X,Y)，以及其边缘分布P(X)的经验分布P^~ (X)如下：



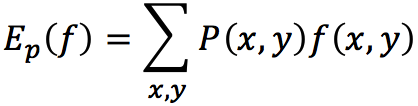
其中，v(X=x, Y=y)表示训练数据集中样本(x, y)出现的频率；v(X=x)表示训练数据集中输入x出现的频率，N是训练数据集的大小。通常可以用特征函数f(x, y)来描述输入x和输出y之间的某一个事实。一般来说，特征函数可以是任意实值函数，下面我们采用一种最简单的二值函数来定义我们的特征函数：



特征函数f(x,y)关于经验分布P^~ (X,Y)的期望值可表示如下：

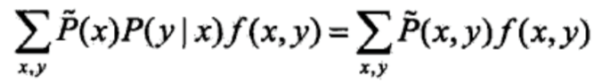


同理可以计算特征函数f(x,y)关于实际联合分布P(X,Y)的数学期望：



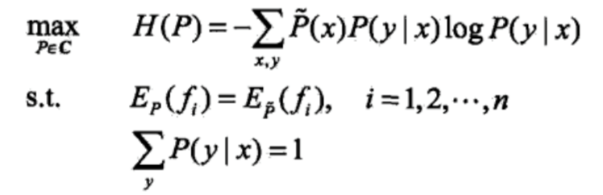


对于我们要求解的条件概率P(Y|X)来说，我们希望这梁这的期望值应该相同，因此有约束条件：



假设在训练数据集中可以得到n个特征函数，那么就能得到n个约束条件。

综上整个最大熵模型的数学模型表示可以表示为：



通过拉格朗日乘子法求解条件极值问题求解。

基于最大熵的文本情感分析只要得到一些训练数据，然后进行迭代，就可以得到所需模型，进行自收敛，方法简单。但是由于最大熵往往只能得到局部最优解而非全局最优解，因此运用该方法进行情感分析准确率有待提高。

* + 1. 基于支持向量机的方法

是一种相对较新的机器学习方法。它通过寻求结构化风险最小以提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。SVM算法的思想是在n维空间里去找到一个决策平面，将数据分为两类，这样最优的超平面被定义为距离决策平面的点到决策平面的距离最大，而这些距离决策平面最近的点称为支持向量。

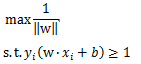
求解SVM算法问题首先要推到SVM模型的目标函数与约束条件，根据SVM算法的定义，首先点x到决策平面的几何距离可表示为：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image005.png

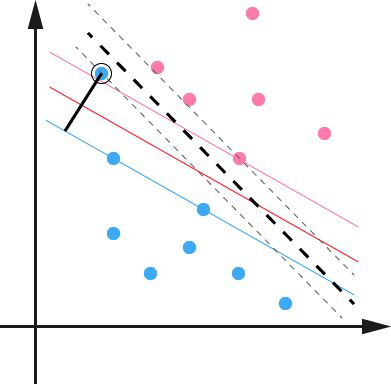
其中 xi 代表第 i 条数据，yi 代表第 i 条数据对应的目标变量的取值，取值有+1 和-1 两种。所以当第 i 条数据被正确分类时，y 取值和 w\*x+b 取值的正负一致，几何间隔为正；当被错误分类时，y 取值和 w\*x+b 取值的正负相反，几何间隔为负。由此可得到其目标函数为max⁡(min⁡(γ\_i))，并遵循约束条件：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image009.png

经过一系列变换之后最终的目标函数与约束条件变成：



但是在真实的数据当中，往往存在一些点并不完全满足上面的约束条件，我们称为特异点，如下图，于是我们引入松弛变量的概念，允许部分点可能会被分到错误的一面。



于是新的目标函数与约束条件就可表示为：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image022.png

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image203.png

在很多情况下往往在当前低维空间下是不能通过线性的超平面分成正负样本的，这个时候就需要通过从低维空间到高维空间的映射使得数据变得线性可分，假设在一个低维空间中可以将训练集分割成两个部分的曲面的公式可表示为：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image037.png

那么通过映射关系：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image038.png

实现在高维空间中的平面

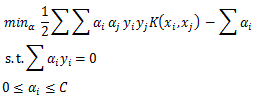
https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image039.png

将数据变得线性可分。

而SVM模型通过引入核函数的概念使得其可以在低维空间中就能完成映射到高维空间的内积运算，无需映射的过程，再结合上面我们求得的目标函数：

https://www.ibm.com/developerworks/cn/analytics/library/machine-learning-hands-on1-svn/image033.png

将其中的包含的自变量内积可以直接转换成核函数，就可以解决线性不可分问题。较为常用的核函数就是高斯核函数，使用核函数之后的木匾函数与约束条件就可以转换成：



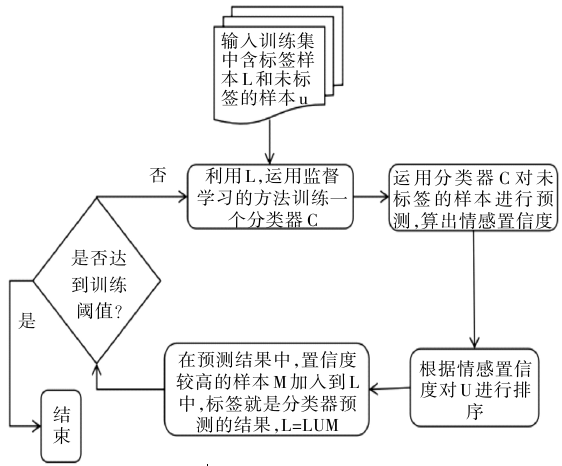
最终就可以通过拉格朗日乘子法求解条件极值问题来得到最优超平面。

基于SVM的文本情感分析方法被认为是最好的情感分析方法，该方法泛化错误率低，计算开销不大，而且对于训练样本较小的文本可以得到很好的情感分析效果，对高维数据的处理效果良好，能够得到较低的错误率，但该方法对参数调节和核函数的选择敏感。

* + 1. 基于半监督的方法

大数据时代，数据的采集尤其是微博等社交媒体文本的采集变得很困难，但是文本的标记却依然相当困难，于是对于有监督学习来说就很难得到足够的数据集。半监督学习的方法可以充分利用大量的未标记样本去改善分类器的性能，对于情感分析任务来说具有很重要的作用。

Self-training的方法是最早提出的一种半监督学习方法，其利用少量标注样本来训练分类器，然后对未标注的文本进行情感倾向预测，找出置信度较高的文本将其预测的结果加入到已标注文本当中去，然后继续迭代，从而不断改善分类器性能，实现文本的情感分类。



由于半监督学习标注的样本数量不够多，导致早期无法很好地训练出准确度高的分类器，很大机率产生错误标注，从而引入噪声。随着噪声的增加，会严重的影响分类器的分类效果。

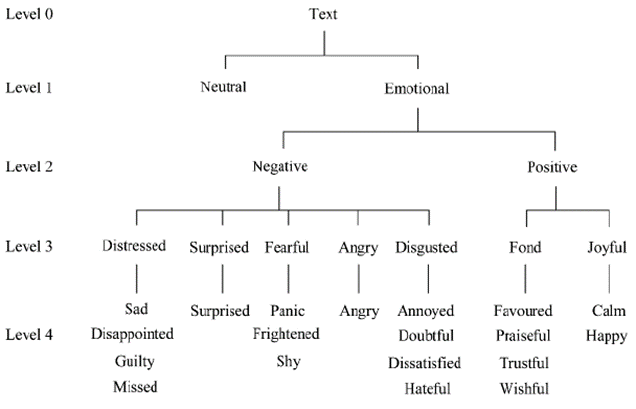
于是就有基于两个方向的半监督学习的改进方法，分别是完全从模型的方向提出更加高效的半监督模型和将模型与语言学相结合分析语料的内在规律并提出相应的解决办法。前者训练阶段不是将所有的己标签样本都用做训练，保留一定比例的标签样本，利用剩余的标签样本训练样本，每次选取后验概率比较低的已标签样本添加进分类器来纠正当前分类器的误差。后者通过对语料的分析，发现文本中往往存在清晰(easy)和模糊(hard)的样本。它通过谱聚类的方法将文本中的情感清晰和情感模糊的样本区分开。对于情感清晰的部分采用聚类的方法将他们分类，通过少量的标注分别给这些样本打标签。对于情感模糊的样本，采用主动学习的方法，将之前打好标签的清晰样本作为训练语料，人工给模糊样本打标签。最后采用集成学习的方法来提高整体情感分类的正确率。

Sun等人将表情符号与一元特征、 二元特征结合起来， 提出了一种面向中文微博的半监督情绪分类方法。该方法利用表情符号对未标记数据进行初始化，将语句中所含表情符号最多的一类标记为该语句的情绪标签；再通过提取语句中词语的一元特征与二元特征，用 SVM 与 朴 素 贝叶斯分类器将微博 中表达的情绪分为7种类别：乐、喜、 悲、 怒、 恐、 恶、 惊。实验表明情绪自动标记的准确率可达到88.7%，情绪分类中朴素贝叶斯分类器要优于SVM其精准率和召回率都超过71%。

* + 1. 其他分类方法

基于有监督的学习方法和基于无监督的学习方法都存在其缺点，于是就有学者考虑将其进行结合的方法，复合分类的方法主要分为两类，其中一类是基于将情绪分类任务分解成无情绪、正负情感、细粒度情感等子任务，再针对不同子任务涉及不同算法的层次情感分类方法。另一类是基于情感词汇本体的多种有监督学习的复合方法。

Xu 等人用层级分类法对中文微博进行了情绪分类，同时还将主成分分析法引入到情绪分类中计算微博中主要情绪的比例。该方法采用4层结构，将情绪细分为19种类别。在新浪微博数据上进行了4层实验，其中第1层只采用平面型文本分类；第2层与第1层不同，采取了层级分类；第3 层在第2 层的基础上加入了词性特征；第4层在第3层的基础上还加入了心理学情绪词典。



* 1. 总结

基于有监督的学习方法将情感分类方法作为特殊的分类任务，训练分类器，实现对情感倾向的预测，但是收集到的数据往往存在大量未标记，标记成本较高，于是半监督学习的方法具有很好的应用前景，但是其也存在一定的缺点，往往会因为预测问题引入噪声或错误分类导致准确率不高。除此之外还会有基于分层的细粒度情感分类，将前两者结合，实现很好的分类。

1. 基于深度学习的方法

近些年来，深度学习火热发展，在自然语言处理与计算机视觉等方面有初中的表现。本节旨在梳理目前主流的用于情感分析中的深度学习方法介绍。

在了解深度学习在情感分析中的应用之前，笔者还是希望读者能够重视情感分析问题的属性。不管是极性的情感分析，还是多情绪状态的情感分析，情感分析都可归结为分类问题，从这个层面来说，理论上可用于分类的神经网络都可用于情感分析的研究。

以下重点介绍用于情感分析主流的两种神经网络：

1. 多层感知器

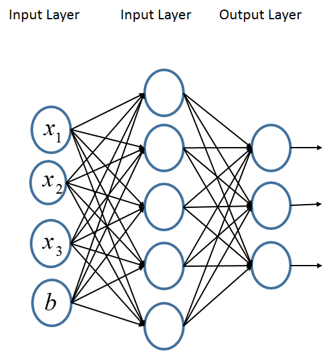
2. 循环神经网络

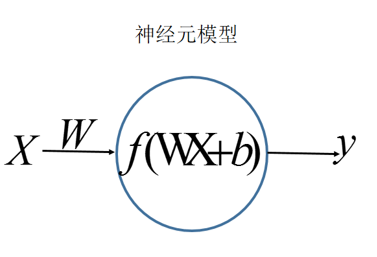
* 1. 多层感知器

多层感知器（Multilayer Perceptron,缩写MLP）是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量。MLP可以被看作是一个有向图，由多个的节点层所组成，每一层都全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元（或称处理单元）。一种被称为反向传播算法的监督学习方法常被用来训练MLP。MLP是感知器的推广，克服了感知器不能对线性不可分数据进行识别的弱点。

多层感知机的一个重要特点就是多层，我们将第一层称之为输入层，最后一层称之有输出层，中间的层称之为隐层。MLP并没有规定隐层的数量，因此可以根据各自的需求选择合适的隐层层数。且对于输出层神经元的个数也没有限制。

MLP神经网络结构模型如下,本文中只涉及了一个隐层，输入只有三个变量[x1,x2,x3]和一个偏置量b，输出层有三个神经元。相比于感知机算法中的神经元模型对其进行了集成。

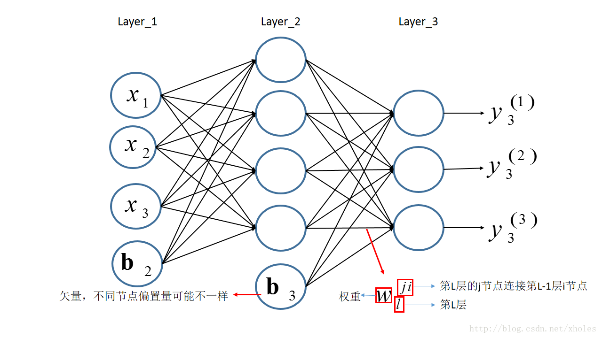




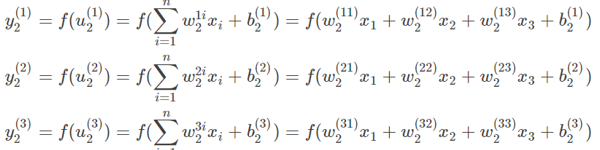
* + 1. 前向传播

前向传播指的是信息从第一层逐渐地向高层进行传递的过程。以下图为例来进行前向传播的过程的分析。

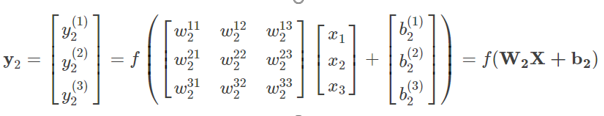
假设第一层为输入层，输入的信息为[x1,x2,x3]。对于层l，用Ll表示该层的所有神经元，其输出为yl,其中第j个节点的输出为y(j)l,该节点的输入为u(j)l，连接第l层与第(l−1)层的权重矩阵为Wl，上一层（第l−1层）的第i个节点到第l层第j个节点的权重为wl(ji)。



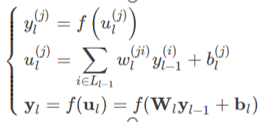
结合之前定义的字母标记，对于第二层的三个神经元的输出则有：



将上述的式子转换为矩阵表达式：



将第二层的前向传播计算过程推广到网络中的任意一层，则:



其中f(⋅)为激活函数，bl(j)为第l层第j个节点的偏置。

* + 1. 反向传播

基本的模型搭建完成后的，训练的时候所做的就是完成模型参数的更新。由于存在多层的网络结构，因此无法直接对中间的隐层利用损失来进行参数更新，但可以利用损失从顶层到底层的反向传播来进行参数的估计。（约定：小写字母—标量，加粗小写字母—向量，大写字母—矩阵）。

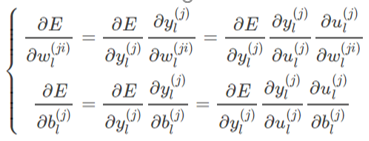
假设多层感知机用于分类，在输出层有多个神经元，每个神经元对应一个标签。输入样本为x=[x1,x2,⋯,xn],其标签为t。

对于层l，用Ll表示该层的所有神经元，其输出为yl,其中第j个节点的输出为yl(j),该节点的输入为ul(j)，连接第l层与第(l−1)层的权重矩阵为Wl，上一层（第l−1层）的第i个节点到第l层第j个节点的权重为wl(ji)。

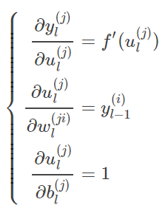
对于网络的最后一层第k层——输出层，现在定义损失函数：



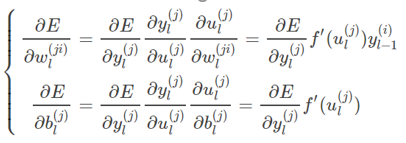
为了极小化损失函数，通过梯度下降来进行推导：



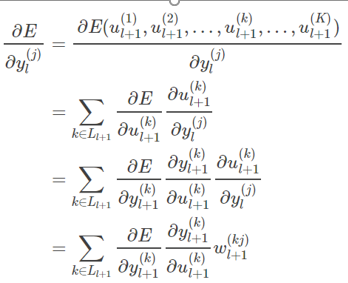
在上式子中，根据之前的定义，很容易得到：



那么则有：



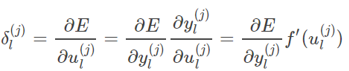
另有，下一层所有结点的输入都与前一层的每个结点输出有关，因此损失函数可以认为是下一层的每个神经元结点输入的函数。那么：



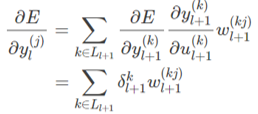
此处定义节点的灵敏度为误差对输入的变化率，即：



那么第l层第j个节点的灵敏度为：



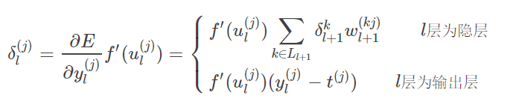
结合灵敏度的定义，则有：



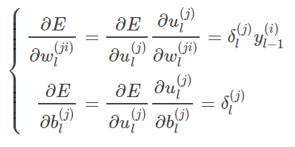
上式两边同时乘上f′(ul(j))，则有



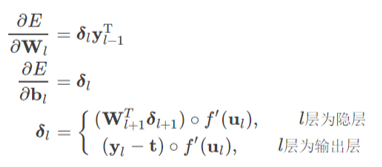
注意到上式中表达的是前后两层的灵敏度关系，而对于最后一层，也就是输出层来说，并不存在后续的一层，因此并不满足上式。但输出层的输出是直接和误差联系的，因此可以用损失函数的定义来直接求取偏导数。那么：



至此，损失函数对各参数的梯度为：

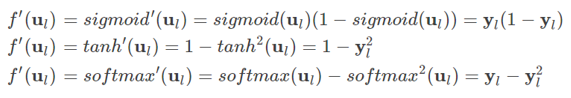


上述的推到都是建立在单个节点的基础上，对于各层所有节点，采用矩阵的方式表示，则上述公式可以写成：

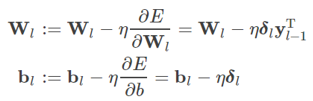


其中运算符∘表示矩阵或者向量中的对应元素相乘。

常见的几个激活函数的导数为：



根据上述公式，可以得到各层参数的更新公式为：



多层感知器具有网络的隐含节点个数选取问题、容易陷入局部极值、学习不够充分。

* 1. 循环神经网络

循环神经网络（Recurrent Neural Network，RNN）是一类具有短期记忆能力的神经网络。在循环神经网络中，神经元不但可以接受其它神经元的信息，也可以接受自身的信息，形成具有环路的网络结构。和前馈神经网络相比，循环神经网络更加符合生物神经网络的结构。循环神经网络已经被广泛应用在语音识别、语言模型以及自然语言生成等任务上。循环神经网络的参数学习可以通过随时间反向传播算法 [Werbos, 1990] 来学习。随时间反向传播算法即按照时间的逆序将错误信息一步步地往前传递。当输入序列比较长时，会存在梯度爆炸和消失问题[Bengio et al., 1994, Hochreiter and Schmidhuber, 1997, Hochreiter et al., 2001]，也称为长期依赖问题。为了解决这个问题，人们对循环神经网络进行了很多的改进，其中最有效的改进方式引入门控机制。

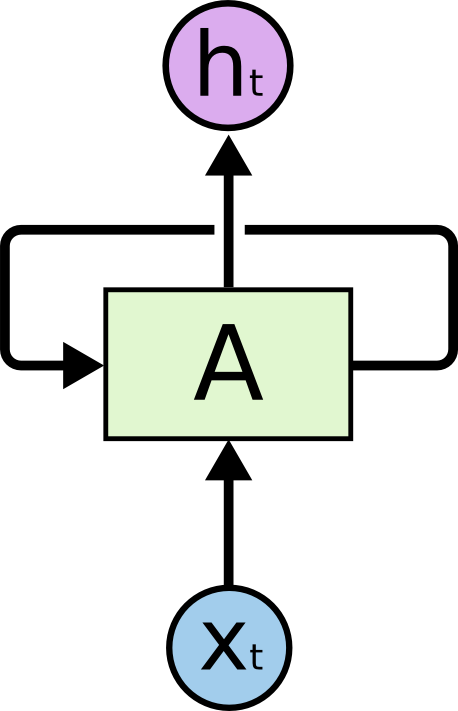
此外，循环神经网络可以很容易地扩展到两种更广义的记忆网络模型：递归神经网络和图网络。

本次课题重点介绍循环神经网络中的长短期记忆网络（LSTM）。

人类并不是每时每刻都从一片空白的大脑开始他们的思考。在你阅读这篇文章时候，你都是基于自己已经拥有的对先前所见词的理解来推断当前词的真实含义。我们不会将所有的东西都全部丢弃，然后用空白的大脑进行思考。我们的思想拥有持久性。

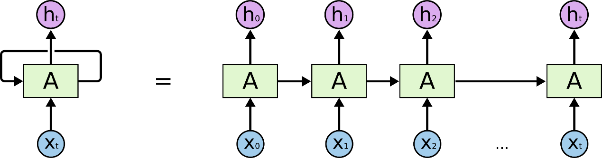
传统的神经网络并不能做到这点，看起来也像是一种巨大的弊端。例如，假设你希望对电影中的每个时间点的时间类型进行分类。传统的神经网络应该很难来处理这个问题——使用电影中先前的事件推断后续的事件。

RNN 解决了这个问题。RNN 是包含循环的网络，允许信息的持久化。



在上面的示例图中，神经网络的模块，A，正在读取某个输入 xi,并输出一个值hi。循环可以使得信息可以从当前步传递到下一步。

这些循环使得 RNN 看起来非常神秘。然而，如果你仔细想想，这样也不比一个正常的神经网络难于理解。RNN 可以被看做是同一神经网络的多次复制，每个神经网络模块会把消息传递给下一个。所以，如果我们将这个循环展开：

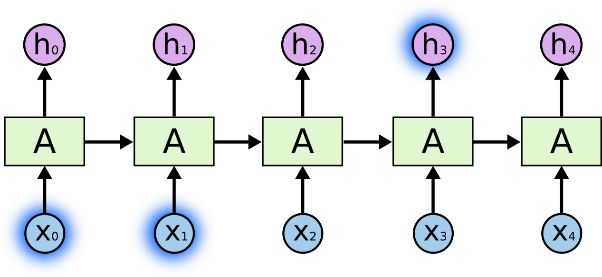


链式的特征揭示了 RNN 本质上是与序列和列表相关的。他们是对于这类数据的最自然的神经网络架构。

* 1. 长期依赖问题

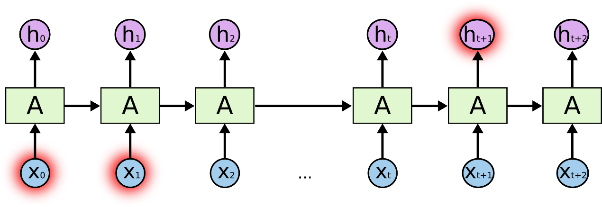
RNN 的关键点之一就是他们可以用来连接先前的信息到当前的任务上，例如使用过去的视频段来推测对当前段的理解。如果 RNN 可以做到这个，他们就变得非常有用。但是真的可以么？答案是，还有很多依赖因素。

有时候，我们仅仅需要知道先前的信息来执行当前的任务。例如，我们有一个语言模型用来基于先前的词来预测下一个词。如果我们试着预测“the clouds are in the sky”最后的词，我们并不需要任何其他的上下文——因此下一个词很显然就应该是sky。在这样的场景中，相关的信息和预测的词位置之间的间隔是非常小的，RNN 可以学会使用先前的信息。



但是同样会有一些更加复杂的场景。假设我们试着去预测“I grew up in France... I speak fluent French”最后的词。当前的信息建议下一个词可能是一种语言的名字，但是如果我们需要弄清楚是什么语言，我们是需要先前提到的离当前位置很远的 France 的上下文的。这说明相关信息和当前预测位置之间的间隔就肯定变得相当的大。

不幸的是，在这个间隔不断增大时，RNN 会丧失学习到连接如此远的信息的能力。



在理论上，RNN 绝对可以处理这样的 长期依赖 问题。人们可以仔细挑选参数来解决这类问题中的最初级形式，但在实践中，RNN 肯定不能够成功学习到这些知识。Bengio, et al. (1994)等人对该问题进行了深入的研究，他们发现一些使训练 RNN 变得非常困难的相当根本的原因。

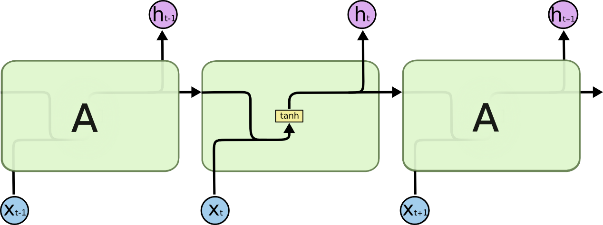
然而，幸运的是，LSTM 并没有这个问题！

* 1. LSTM
     1. LSTM网络

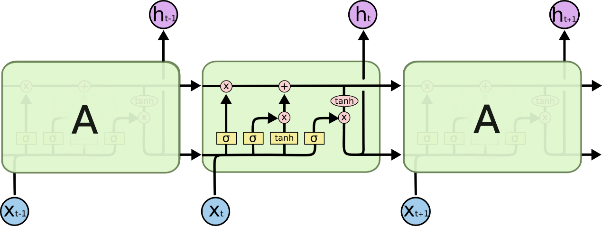
Long Short Term网络，一般就叫做 LSTM，是一种RNN特殊的类型，可以学习长期依赖信息。LSTM由Hochreiter & Schmidhuber (1997)提出，并在近期被Alex Graves进行了改良和推广。在很多问题，LSTM都取得相当巨大的成功，并得到了广泛的使用。

LSTM通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM的默认行为，而非需要付出很大代价才能获得的能力！

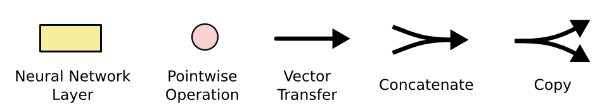
所有 RNN 都具有一种重复神经网络模块的链式的形式。在标准的 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。



LSTM同样是这样的结构，但是重复的模块拥有一个不同的结构。不同于 单一神经网络层，这里是有四个，以一种非常特殊的方式进行交互。



不必担心这里的细节。我们会一步一步地剖析 LSTM解析图。现在，我们先来熟悉一下图中使用的各种元素的图标。

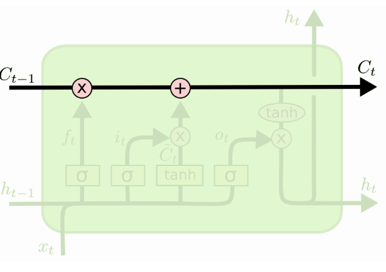


在上面的图例中，每一条黑线传输着一整个向量，从一个节点的输出到其他节点的输入。粉色的圈代表按位 pointwise的操作，诸如向量的和，而黄色的矩阵就是学习到的神经网络层。合在一起的线表示向量的连接，分开的线表示内容被复制，然后分发到不同的位置。

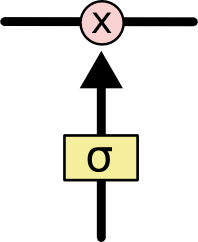
* + 1. LSTM的核心思想

LSTM的关键就是细胞状态，水平线在图上方贯穿运行。

细胞状态类似于传送带。直接在整个链上运行，只有一些少量的线性交互。信息在上面流传保持不变会很容易。



LSTM有通过精心设计的称作为“门”的结构来去除或者增加信息到细胞状态的能力。门是一种让信息选择式通过的方法。他们包含一个 sigmoid 神经网络层和一个按位的乘法操作。



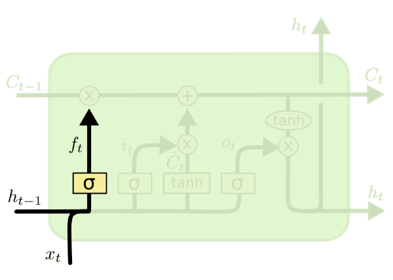
Sigmoid 层输出 0 到 1之间的数值，描述每个部分有多少量可以通过。0代表“不许任何量通过”，1 就指“允许任意量通过”！

LSTM 拥有三个门，来保护和控制细胞状态。

* + 1. 逐步理解LSTM

在我们LSTM中的第一步是决定我们会从细胞状态中丢弃什么信息。这个决定通过一个称为忘记门层完成。该门会读取 ht-1和xt,输出一个在0到1之间的数值给每个在细胞状态Ct-1中的数字。1表示“完全保留”,0表示“完全舍弃”。

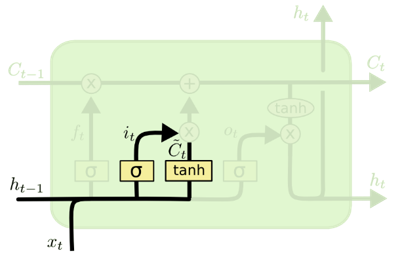
让我们回到语言模型的例子中来基于已经看到的预测下一个词。在这个问题中，细胞状态可能包含当前主语的性别，因此正确的代词可以被选择出来。当我们看到新的主语，我们希望忘记旧的主语。

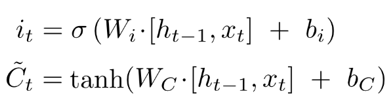




下一步是确定什么样的新信息被存放在细胞状态中。这里包含两个部分。第一，sigmoid层称“输入门层”决定什么值我们将要更新。然后，一个tanh层创建一个新的候选值向量，Ct，会被加入到状态中。下一步，我们会讲这两个信息来产生对状态的更新。

在我们语言模型的例子中，我们希望增加新的主语的性别到细胞状态中，来替代旧的需要忘记的主语。

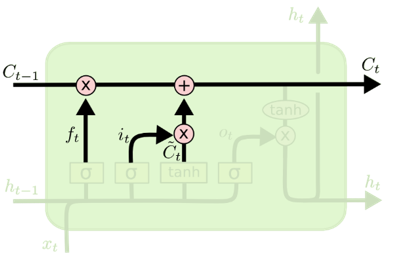




现在是更新旧细胞状态的时间了，Ct-1更新为Ct。前面的步骤已经决定了将会做什么，我们现在就是实际去完成。

我们把旧状态与ft相乘，丢弃掉我们确定需要丢弃的信息。接着加上it\*Ct。这就是新的候选值，根据我们决定更新每个状态的程度进行变化。

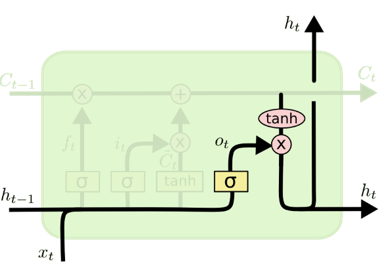
在语言模型的例子中，这就是我们实际根据前面确定的目标，丢弃旧代词的性别信息并添加新的信息的地方。

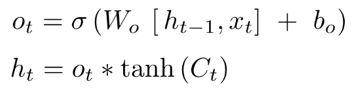




最终，我们需要确定输出什么值。这个输出将会基于我们的细胞状态，但是也是一个过滤后的版本。首先，我们运行一个sigmoid层来确定细胞状态的哪个部分将输出出去。接着，我们把细胞状态通过tanh进行处理（得到一个在-1到1之间的值）并将它和sigmoid门的输出相乘，最终我们仅仅会输出我们确定输出的那部分。

在语言模型的例子中，因为他就看到了一个代词，可能需要输出与一个 动词 相关的信息。例如，可能输出是否代词是单数还是负数，这样如果是动词的话，我们也知道动词需要进行的词形变化。





* 1. 结论

刚开始，我提到通过RNN得到重要的结果。本质上所有这些都可以使用 LSTM完成。对于大多数任务确实展示了更好的性能！

由于LSTM一般是通过一系列的方程表示的，使得LSTM有一点令人费解。然而本文中一步一步地解释让这种困惑消除了不少。

LSTM是我们在RNN中获得的重要成功。很自然地，我们也会考虑：哪里会有更加重大的突破呢？在研究人员间普遍的观点是：“Yes! 下一步已经有了——那就是注意力！”这个想法是让RNN的每一步都从更加大的信息集中挑选信息。例如，如果你使用RNN来产生一个图片的描述，可能会选择图片的一个部分，根据这部分信息来产生输出的词。实际上，Xu, et al.(2015)已经这么做了——如果你希望深入探索注意力可能这就是一个有趣的起点！