1. CRF算法概述

条件随机场（conditionl random fields,CRFs）由J.Lafferty等人（2001）提出，近几年来在自然语言处理和图像处理等领域得到广泛的应用。CRF是用来标注和划分序列结构数据的概率化结构模型。言下之意就是对于给定的输出标识Y和观察序列X，条件随机场通过定义条件概率P(Y|X)，而不是联合概率分布P(Y，X)来描述模型。CRF也可看作是一个无向图模型或者马尔可夫随机场（Markov random field）。

（1）条件随机场定义

设*X*与*Y*是随机变量，*P*(*Y*|*X*)是在给定*X*的条件下*Y*的条件概率分布，若随机变量*Y*构成一个由无向图*G*=(*V*,*E*)表示的马尔可夫随机场，即

*P*(*Yv*|*X*，*Yw*,*w*)= *P*(*Yv*|*X*,*Yw*,*w*)

对任意结点*v*成立，则称条件概率分布*P*(*Y*|*X*)为条件随机场。式中*w*表示在图*G*=(*V*,*E*)中与结点有边连接的所有结点*w*，*w*表示结点以外的所有结点，*Yv*，*Yu* 与*Yw*为结点，与对应的随机变量。

在定义中并没有要求*X*和*Y*具有相同的结构。现实中，一般假设*X*和*Y*有相同的图结构。本书主要考虑无向图为如图1-2与图1-3所示的线性链的请况，即

G=（V={1,2,…,*n*},E={(*i*，*i*+1)}），*i=*1,2,…,*n*-1

在此情况下，*X*=（*X*1，*X*2,…,*X*n）,*Y*=（*Y*1，*Y*2,…,*Y*n）,最大团是相邻两个结点的集合。线性链条件随机场有下面定义。

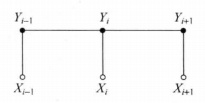
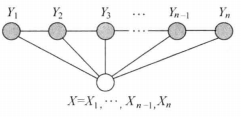


图1-3 线性链条件随机场 图1-4有相同图结构的线性链条件随机场

（2）线性链条件随机场定义

设*X*=（*X*1，*X*2,…,*X*n）,*Y*=（*Y*1，*Y*2,…,*Y*n）均为线性链表示的随机变量序列，若在给定随机变量序列*X*的条件下，随机变量序列*Y*的条件概率分布*P*(*Y*|*X*)构成条件随机场，即满足马尔可夫性

*P*(*Yi*|*X*，*Y*1,..., *Yi*+1,*Yi*+1,…,*Y*n)= *P*(*Yi|X*,*Yi*-1, *Yi*+1)

*i*=1,2,…,*n*(在*i*=1和*n*时只考虑单边)

则称*P*(*Y*|*X*)为线性链条件随机场。在标注问题中，*X*表示输入观测序列，*Y*表示对应的输出标记序列或状态序列。

条件随机场模型也需要解决三个基本问题：特征的选择、参数训练和解码。其中，参数训练过程可在训练数据集上基于对数似然函数的最大化进行，【具体算法参照文献。。。】

相对于HMM，CRF的主要优点在于它的条件随机性，只需要考虑当前已经出现的观测状态的特性，没有独立的严格要求，对于整个序列内部的信息和外部观测信息均可有效利用，避免了MEMM和其他针对线性序列模型的条件马尔科夫模型会出现的标识偏置问题。CRF具有MEMM的一切优点，两者的关键区别在于，MEMM使用每一个状态的指数模型来计算给定前一个状态下当前状态的条件概率，而CRF用单个指数模型来计算给定观察序列与整个标记序列的联合概率。因此，不同状态的不同特征权重可以相互替换。

条件随机场的一个重要应用是在文本分析领域。针对不同的任务，提出了各种模型，包括词性标注、识别命名实体（人名、地名、组织名等）和抽取结构化信息（如从参考文献列表中抽取出版物的标题、作者、杂志、年份等）。在这些模型中，大部分模型都共享一个相似的结构：文档中的每个单词（也可能是短语）都对应着代表每个单词类别的一个目标变量。每个目标变量都与一个特征向量相关，用于刻画与目标特性相关的性质。这些方法在文本分析中非常流行，一方面因为网络结构非常适合这个领域，另一方面因为它们在自然语言处理问题的一个很大范围内取得了最好的效果。

1. Word2vec词义表示模型

（1）Word2vec的训练模型

Word2vec是Mikolov等所提出模型的一个实现，可以用来快速有效地训练词向量。Word2vec包含了两种训练模型，分别是CBOW和Skip\_gram，如图２所示。

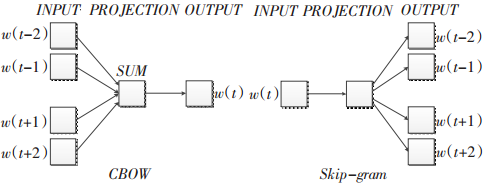
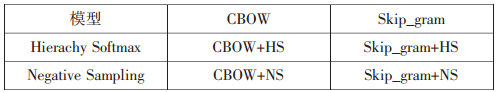


图1-5 CBOW和Skip\_gram模型

从图２可以看出，CBOW和Skip\_gram模型均包含输入层、投影层和输出层。其中，CBOW模型通过上下文来预测当前词，Skip\_gram模型则通过当前词来预测其上下文。Word2vec提供了两套优化方法来提高词向量的训练效率，分别是Hierachy Softmax和Negative Sampling。将训练模型和优化方法进行组合可得到４种训练词向量的框架，如表１所示。

表1 Word2vec词向量训练框架



（2）Word2vec训练词向量原理

CBOW+HS和Skip\_gram+HS:从Word2vec的源码中可以归纳出CBOW+HS和Skip\_gram+HS模型的训练框架，如图1-6所示。两种训练框架的共同点和区别，如表2所示。

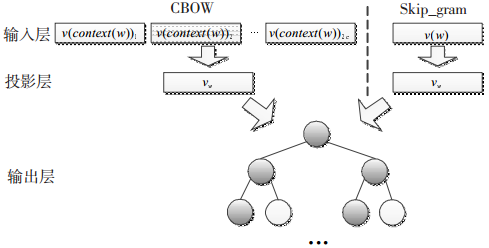
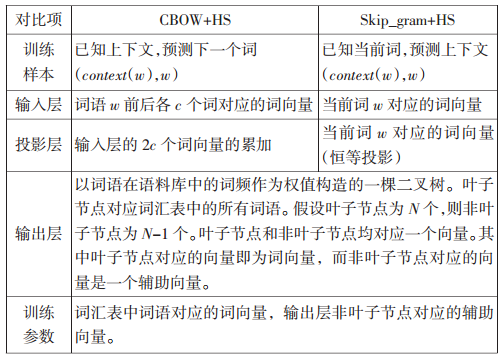


图1-6 CBOW+HS和Skip\_gram+HS模型的训练框架

表2 CBOW+HS和Skip\_gram+HS训练框架对比



在模型的训练过程中，梯度是训练参数更新的依据。为了获得梯度公式，需要构造出训练模型的目标函数。 在CBOW+HS框架的训练中，给定一个训练样本（context（w），w），则在输出层从根节点到词*w*的路径上，对于每一个非叶子节点均对应一个辅助向量j和一次二分类且左右两边各对应一个哈夫曼编码j。其中，j表示从根节点到叶子节点w路径上非叶子节点的编号。对于j，规定向左的分支对应１，向右的分支对应０。这样，通过若干二分类后，最终可到达叶子节点w。可将非叶子节点向左的分支定义为负类，将向右的分支定义为正类。通过逻辑回归知识，可以得到一个二分类被分成正类的概率为,被分成负类的概率为。其中，w是 context （w）中包含的词所对应词向量的累加和，而为非叶子节点对应的辅助向量。沿着从根节点到叶子节点 w 的路径将每个非叶子节点的二分类概率相乘即为 p（w｜context（w））。

对于Skip\_gram+HS，给定一个训练样本（w，context（w）），其中 context（w）包含2c个词。可以将通过w预测context（w）的问题，即p（context（w）｜w）转换成2c个通过w预测下一个词为u的问题p（u｜w），其中u∈context（w）。p（u｜w）可以利用CBOW+HS框架的思路来解决，即将u视为叶子节点。不同的是，在CBOW+HS框架中，w指的是context（w）中所有词对应词向量的累加，而在Skip\_gram+HS中指的是w对应的词向量。Skip\_gram+HS框架的目标函数为。

CBOW+NS和Skip\_gram+NS：CBOW+HS和Skip\_gram+HS框架输出层的哈夫曼树构造过程相对复杂。 所以，CBOW+NS和Skip\_gram+NS框架采用了一种替代的方法， 即采用相对简单的负取样来提高词向量训练的速度。

对于CBOW+NS框架，已知样本（context（w），w），则w为一个正样本，而词汇表中的其他词为负样本。可以采用不同的负采样算法来选择负样本。在确定关于context（w）的一个非空的负样本集合NEG（w）后，对于词汇表中的任何词w′，若w′＝w，给其为１的标签，否则给其为０的标签。这样，正样本的标签为１，负样本的标签为０。这里的标签可类比成CBOW+HS框架输出层哈夫曼树中非叶子节点的左右二分类编码。优化的目标函数可表示为。

对于Skip\_gram+NS框架，已知样本（w，context（w））。则对于一个上下文词w′，当w′＝w，给其为１的标签，否则给其为０的标签。这里的标签同样可类比成Skip\_gram+HS框架输出层哈夫曼树中非叶子节点的左右二分类编码。则优化的目标函数为。

1. Seq2Seq生成算法模型

Seq2Seq主要思路是先将输入序列编码成一个向量，之后通过提取向量中的信息解析输出序列。Seq2Seq结构由编码器和解码器组成。其中，编码器用于编码输入序列，而解码器用于解析输出序列。本文中的自动问答系统模型的网络结构是基于Seq2Seq的一种基础结构组成，其结构如图1-7所示。

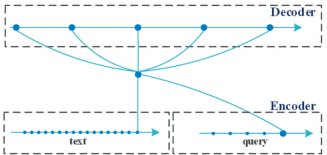


图1-7 模型结构图

模型分为两部分，一部分是 encoder，将输入序列进行编码表示，另一部分是 decoder，以 encoder 的表示作为 condition，将目标序列解码出来。由于文本通常是一个变长的序列，所以通常用 RNN、LSTM、GRU 以及他们的变种模型对文本进行建模。