关于命名实体识别的生成式对抗网络的研究

冯建周,马祥聪,刘亚坤,宋沙沙

(燕山大学 信息科学与工程学院,河北 秦皇岛 066004)
(燕山大学 河北省软件工程重点实验室,河北 秦皇岛 066004)
E-mail: fjzwxh@ysu.edu.cn

摘 要: 本文结合条件生成式对抗网络(CGAN)和改进的 Wasserstein 生成式对抗网络(WGAN-GP),提出一种适合于命名实体识别任务的条件 Wasserstein 生成式对抗网络模型(CWGAN). 该模型借鉴 CGAN 以文本描述为条件的图像概率分布的思想,来完成以句子序列为条件获得标注序列概率分布的任务. 该模型的生成器和判别器都采用 BiLSTM 结构,不同的是生成器生成命名实体标签的概率分布,判别器则为生成器的生成质量打分并反馈给生成器,生成器根据反馈更新梯度从而提升生成标签概率的质量. 另外,CWGAN 采用梯度惩罚的方法来保证梯度在反向传播的过程中保持平稳,通过拉近真实样本分布和生成样本之间的 Wasserstein 距离,优化目标函数. 最后通过实验验证了该方法的可行性和优越性.

关 键 词: 命名实体识别; 生成式对抗网络; BiLSTM; Wasserstein 距离; CWGAN
中图分类号: TP311 **文献标识码**: A **文 章 编 号**: 1000-1220(2019)06-1191-06

Research on Generative Adversarial Networks of Named Entity Recognition

FENG Jian-zhou, MA Xiang-cong, LIU Ya-kun, SONG Sha-sha
(Yanshan University College of Information Science and Engineering, Qinhuangdao 066004, China)
(Yanshan University Key Laboratory of Hebei Software Engineering, Qinhuangdao 066004, China)

Abstract: This paper proposed a Generative Adversarial Nets suitable for the task of named entity recognition named Conditional Wasserstein Generative Adversarial Nets(CWGAN), inspired from Conditional GAN and improved Wasserstein GAN. Relative to the image probability distribution conditioned on textual description in CGAN, CWGAN obtains the NER label sequence probability distribution conditioned on sentence sequences. Both the generator and the discriminator use a bidirectional LSTM network. The difference is that the generator generates the probability distribution of the named entity tags, and the discriminator scores the generation quality of the generator and feeds it back to the generator. The generator updates the gradient according to the feedback to improve the quality of the probability of generating tags. In addition, this paper use gradient penalty in improved Wasserstein GAN to ensure that the gradient remains stable during backward propagation. Meanwhile, this paper use the mean which decrease the Wasserstein distance between real sample distribution and generate sample ensure that the target function is optimized. Experiments show that the CWGAN model we proposed is effective in the task of named entity recognition. Finally, the feasibility and superiority of the method are verified by experiments.

Key words: named entity recognition; generative adversarial networks; bidirectional LSTM; wasserstein distance; conditional wasserstein generative adversarial nets(CWGAN)

1 引 言

互联网的快速发展使网络信息呈爆发式增长,同时网络信息的形式也变得越来越多样化,这给用户有效利用网络信息资源带来了很大的不便. 面对网络信息爆发式增长带来的挑战,信息抽取技术逐渐发展起来. 信息抽取是指从大规模的无结构文本中提取出用户真正感兴趣的信息,并以结构化或半结构化的形式存储或输出[1].

信息抽取技术起源于 20 世纪 70 年代早期对自然语言处理(Natural Language Processing, NLP)的研究,而后从 20 世纪 80 年代中期开始蓬勃发展起来,这得益于消息理解会议(Message Understanding Conference, MUC)[1] 的推动. 继 MUC 之后,自动内容抽取(Automatic Content Extraction, ACE)[2] 评测会议也对信息抽取技术的发展起着关键性的作用.

根据 ACE 的划分,信息抽取主要包括 4 个方面的研究:命名实体识别、指代消解、实体关系抽取和事件抽取. 其中,命名实体识别(Named Entity Recognition, NER)是这些任务中最关键的部分. 这是因为命名实体识别是 NLP 领域中一些复杂任务(如机器翻译、问答系统、信息检索等)的基础. 同时命名实体识别又是实体关系抽取的基础. 例如,在机器翻译中,

[1] MUC. http://en.wikipedia.org.wiki/Message_Understanding_Conference.
[2] ACE. http://en.wikipedia.org.wiki/Automatic_Content_Extraction.

收稿日期: 2018-08-28 收修改稿日期: 2018-09-26 基金项目: 国家自然科学基金青年基金项目(61602401)资助; 河北省高等学校科学技术研究青年基金项目(QN2018074)资助. 作者简介: **冯建周**(通讯作者),男,1978 年生,博士,副教授,CCF 高级会员,研究方向为知识图谱、语义 web; **马祥聪**,男,1994 年生,硕士,研究方向为知识库补全; **刘亚坤**,男,1997 年生,研究方向为命名实体识别、实体关系抽取; **宋沙沙**,女,1992 年生,硕士,研究方向为命名实体识别、实体关系抽取.

只有将目标句子中的实体准确地识别出来并知道实体之间的语义关系才能够准确的翻译目标句子. 在问答系统中,系统只有从用户的提问中准确地识别出实体类型以及实体之间的关系才能更好地为用户解答.

命名实体识别任务最初是在 MUC-6 上被提出的,它的主要任务是识别出自然语言文本中的各种短语并加以归类. 它有两个关键的任务: 一是要识别出文本中是否有命名实体,二是要判断出命名实体具体所指的目标类型. 命名实体的领域相关性很强; 数量巨大,收录非常困难,没有通用化的字典可供查询; 表达形式多样,可能采用缩写等其他变化的方式,影响判别准确率等等,这些都给命名实体识别任务增加了难度.

2 相关工作

命名实体识别是 NLP 领域中一些复杂任务的基础,因此一直以来都是 NLP 领域中的研究热点. 现有的命名实体识别研究方法有基于规则的方法,基于传统机器学习的方法(又叫统计的方法) ,以及近年来流行的基于深度学习的方法.

基于规则的方法[2] 由于手工构造规则,系统能够达到较好的性能,但构造规则时太依赖于专业领域知识,费时费力且系统的可移植性较差.

基于机器学习的方法中,命名实体识别被看作是序列标注问题,传统的机器学习方法有许多适用于序列标注问题的模型. Borthwick 等人[3] 利用最大熵马尔科夫模型和额外知识集提高了 NER 的准确性. Lafferty 等人[4] 提出条件随机场用于模式识别任务. Zhou 等人[5] 提出使用四种不同特征来提高隐马尔可夫模型在 NER 任务上的性能. McCallum A[6] 提出使用更丰富,更高阶的马尔科夫模型的特征感应法和维特比法用于 NER 任务. 除了基于有监督的机器学习方法,机器学习的半监督和无监督的学习方法也可以用于 NER 任务. 在 NER 方面,主要的半监督学习方法是“bootstrapping”方法[7,8]. 李丽双[9] 利用半监督 SVM 模型与 CRF 模型进行组合的方法,实现了将多分类器组合与字典匹配运用到命名实体识别中,提高了试验效果. 此外,还有一些无监督的开放信息抽取系统,如华盛顿大学的 TxtRunner[10]、ReVerb[11] 等系统,斯坦福大学的 Stanford OpenIE[3] 等是开放信息抽取中的典型工作.

近年来,随着深度学习算法的普及,很多学者开始将深度学习算法应用在 NER 领域,而且已经取得了卓越的效果. Athavale V[12] 提出的 BiLSTM 模型采用了双向长短时记忆(Bi-Long-Short-Term Memory,BiLSTM) 网络,通过 BiLSTM 网络将上下文结合起来,进行 NER 的训练,取得了良好的效果. Huang Z[13] 和 Lample G[14] 采用 BiLSTM 与 CRF(Conditional random field algorithm) 结合的方法,进行命名实体识别的实验,不但能充分利用上下文的信息,又能考虑到句子的语义规则信息,从而取得了比单纯 BiLSTM 更好的效果. Chiu 等人[15] 使用 BiLSTM + CNN 模型来获取更多的特征,在输入层,将词向量和词特征进行结合,然后利用 CNN 进行特征抽取,最后,通过 BiLSTM 进行训练,从而提高了效果. Rei 等人[16] 在 RNN-CRF 模型结构基础上,重点改进了词向量与字符向量的拼接,采用 CRF 作为输出层,并以预测的标签作为条件,使用注意力机制(Attention)[4] 将原始的字符向量和词向量拼接改进成权重求和,使用两层传统神经网络隐层来学习 Attention 的权值,这样就使得模型可以动态地利用词向量和字符向量信息. 深度学习的方法在 NER 领域虽然取得了很好的效果,但是仍然存在很大的改进空间,比如超参数的选择仍然依赖经验,优化过程过早收敛等情况.

2014 年,Ian Goodfellow[17] 提出了生成式对抗网络,即 GAN(Generative Adversarial Networks) 模型. 最初,GAN 模型是用于生成图像这样的连续数据的,并不能直接用来生成离散数据. 而当离散数据做微小改变时,在映射空间中也许根本就没有对应意义的序列,所以当 GAN 处理 NLP 这种离散数据的任务时,容易出现梯度消失的问题. 此外,GAN 无法判断目前生成的某一部分序列的质量,因为它只能给生成的完整序列打分.

但是,这些问题近两年已经有所突破. 于澜涛等人[18] 提出的 SeqGAN(Sequence Generative Adversarial Nets) 模型,通过执行强化学习中的策略梯度解决了原始 GAN 在序列标注问题中无法为生成器提供梯度的问题. SeqGAN 中的奖励信号仍来自判别器对完整序列的判断,只不过它使用蒙特卡洛搜索返回中间状态的动作步骤来实现为部分序列打分. Arjovsky M[19] 提出了 WGAN 模型来解决 NLP 领域的梯度消失问题. 该论文给出了 GAN 训练效果不稳定的原因,并利用 wassertein 距离进行了解决,同时解决了 GAN 的模式崩溃的问题. Mirza M[20] 提出的 CGAN 模型针对 NLP 领域以往的 GAN 不能生成特定属性的问题,进行了相关改进,它将特定属性融入到生成器和判别器当中,从而解决了 GAN 不能生成特定属性的缺点. Gulrajani I[21] 在 WGAN 的基础之上提出了 WGAN-GP,通过采用 lipschitz 连续性限制的方法,解决了训练梯度消失或者梯度爆炸的问题,同时,提高了收敛速度.

相关研究工作表明,GAN 可以在 NLP 任务上有杰出表现. 但在 NER 方面,GAN 还没有相应的研究. 因此,本文将 CGAN 和 WGAN-GP 两者的优点结合,提出一个适合于命名实体识别任务的条件 Wasserstein 生成式对抗网络(Conditional Wasserstein Generative Adversarial Nets,CWGAN) .

3 基于 CWGAN 的命名实体识别

命名实体识别任务一般被看做序列标注问题. 因此,本文将未标注的句子作为条件,构建 CWGAN 模型,完成命名实体识别任务. 在对抗学习中本文将命名实体识别任务描述如下: 给定一个未标注的由一系列单词组成的句子 $X = \{x_1, x_2, \cdots, x_n\}$,并以此作为 CWGAN 模型的条件,生成器模型通过条件生成句子的标注序列,判别器模型给生成的标注序列打分,并为生成器模型提供反馈指导生成器模型训练,最终训练好的生成器模型能够生成质量较高的命名实体标签 $Y = \{y_1, y_2, \cdots, y_n\}$,其中 x_i 代表单词,y_i 代表其对应的生成标签.

3.1 用于命名实体识别的 CWGAN 模型的设计思路

本小节介绍用于命名实体识别的 CWGAN 模型的设计

[3] Stanford Open Information Extraction. https://nlp.stanford.edu/software/openie.html.

[4] http://blog.heuritech.com/2016/01/20/attention-mechanism/

思路. 如图 1 所示,模型分为两部分: 生成器模型(G) 和判别器模型(D) .

生成器模型(G) 定义了在给定句子的情况下生成该句子对应的命名实体标签序列的策略. 本文的生成器模型使用的是一个 BiLSTM 网络. 将句子序列 X 输入到生成器中(句子序列是未标注的) ,将未标注的句子作为条件信息,用于生成标注,通过 BiLSTM 网络得到每个单词的上下文信息表示,通过全连接层以及 softmax 层得到每个单词在各个命名实体标签上的概率. 判别器模型(D) 使用的是一个 BiLSTM 网络. 将句子序列 X 及其对应标注标签序列 L(专家预先标注的标签) 连接作为正实例输入到 D,同时,将句子序列 X 与 G 生成的标签序列连接起来作为负实例输入到 BiLSTM 网络,以专家标注的标签为参照,为生成的标签序列的打分,将每个词的生成标签得分进行求和,返回句子中每个词的标签的总得分,最后句子的得分是句子中每个词的得分的均值. 由于得分均值是通过每个词和标注组合的得分加和而来,所以在反向传导过程中,D 能针对 G 每一步的输出进行反馈. D 给 G 的反馈是针对于句子中每个词的,这样做相对于直接对整个标签序列和句子序列进行判别,返回的信息更多,更有助于 G 的优化.

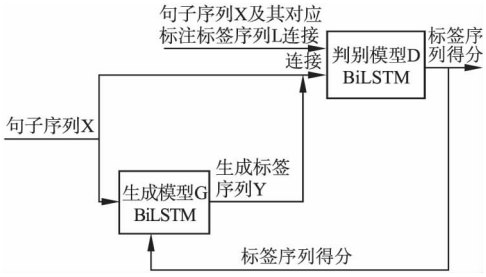

图 1　CWGAN 整体框架

Fig. 1　Whole frame diagram CWGAN

3.2　输入表示

NLP 任务中通常要把文本转换成分布式表示,本文采用词向量的方式来表示文本中的单词,使用词向量的目的是将句子中的每个单词映射成 K 维实值向量. 例如,给定一个句子 $X=\{x_1,x_2,\cdots,x_n\}$,通过映射词向量矩阵 $E\in\mathbf{R}^{|V|\times d_w}$ 将每个单词 $\boldsymbol{x}_i$ 表示为 $\boldsymbol{d}_w$ 维实值向量,$\boldsymbol{V}$ 是词表的大小(词向量训练语料中的词的数目) .

3.3　生成器模型

生成器模型使用的是一个普通的 BiLSTM 网络. 对于给定的包含 n 个单词的句子 $X=(x_1,x_2,\cdots,x_n)$,将每个单词表示为 d_w 维实值向量,用 BiLSTM 结构分别计算句子中每个单词的左上下文表示 $\overrightarrow{h}_t$ 和右上下文表示 $\overleftarrow{h}_t$. 前者称为前向 LSTM,后者称为后向 LSTM,这是两个不同参数的网络. 连接其左右上下文表示 $h_t=[\overrightarrow{h}_t;\overleftarrow{h}_t]$,$h_t$ 即为单词 x_t 的上下文信息.

如图 2 所示,得到单词的上下文表示后,通过一个全连接层并将结果传送给 softmax 层得到每个单词在各类标签上的条件概率,计算方法如下:

$$p(y|x_i;\theta)=\frac{\exp(o_y)}{\sum_{k=1}^{n_m}\exp(o_k)},1\leqslant i\leqslant T \tag{1}$$

公式(1) 表示在参数 θ 下,单词 x_i 的标签归为 y 的概率,其中,n_m 是标签种类个数,o 是全连接层后的输出,其计算方法如下:

$$o_i^y=W'h_i+b' \quad ,\quad 1\leqslant i\leqslant T \tag{2}$$

其中,W′代表全连接层的权重矩阵,h_i 代表单词 x_i 的上下文信息,b'为偏置向量.

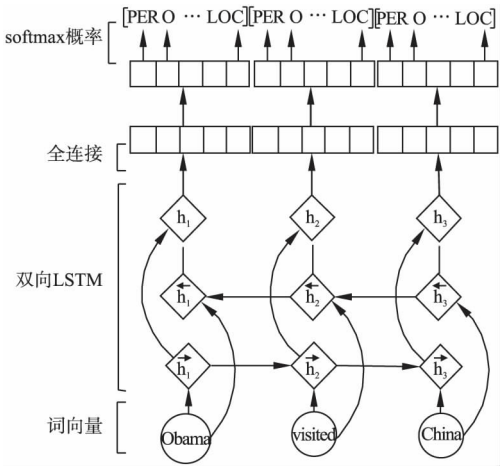

图 2　生成器的 BiLSTM 处理过程

Fig. 2　Bidirectional LSTM processing of generator

3.4　判别器模型

本文提出的 CWGAN 模型中的判别器模型也是 BiLSTM 网络. 使用 BiLSTM,可以为每个单词的标签单独打分. 由于句子的长度是不固定的,通过填充将句子转换为具有固定长度 T 的序列,该长度是生成器模型的输出设置的最大长度.

判别器的 BiLSTM 的输入有两种,一种是句子序列 $x_1,\cdots,x_T$ 和对应的标注标签序列 $l_1,\cdots,l_T$ 的连接 $[X;L]$,另一种句子序列 $x_1,\cdots,x_T$ 和生成器生成的标签序列 $y_1,\cdots,y_T$ 的连接 $[X;Y]$(两个句子序列相同,标签不同) . 其中 X,L,Y 分别为句子序列矩阵 $X_{1:T}$ 和生成标签矩阵 $Y_{1:T}$ 以及标注标签矩阵序列 $L_{1:T}$,它们分别建立为:

$$X_{1:T}=x_1;x_2;\cdots;x_T$$
$$Y_{1:T}=y_1;y_2;\cdots;y_T$$
$$L_{1:T}=l_1;l_2;\cdots;l_T$$

其中,x_t,y_t,l_t 是 k 维词向量表示,分号是连接运算符,且在 $[X;L]$ 和 $[X;Y]$ 中为行连接,在 $X_{1:T}$、$Y_{1:T}$ 以及 $L_{1:T}$ 中为列连接,即,若标签序列维度为 n_1,则连接后的 $[X;L]$ 和 $[X;Y]$ 的维度为 $T\times(d_w+n_1)$.

将连接后的向量矩阵作为 BiLSTM 的输入,BiLSTM 结构会分别计算句子中每个单词的左上下文表示 $\overrightarrow{h'}_t$ 和右上下文表示 $\overleftarrow{h'}_t$(计算过程同 3.3 小节) 然后连接其左右上下文表示 $h'_t=[\overrightarrow{h'}_t;\overleftarrow{h'}_t]$,$h'_t$ 则包含了单词序列与标签序列连接后的序列的上下文信息(与之对应的 3.3 节的 h,包含的是未标注的单词序列及其未标注的上下文信息) . 如图 3 所示,得到单词序列与标签序列连接后序列的上下文表示后,通过一个全连接层计算每个单词及其标签的连接序列的得分,计算方法如下:

$$s_i=W^*h'_i+b^* \quad ,1\leqslant i\leqslant T \tag{3}$$

其中，W^* 代表全连接层的权重矩阵，h'_i 为单词及其标签连接序列的上下文信息，b^* 为偏置向量. 最后，将每个单词及其标签连接序列的得分相加，得到整个句子标注后的得分. 当训练判别器的时候，它的输入是句子序列 $x_1,\cdots,x_T$ 和对应的标注标签序列 $l_1,\cdots,l_T$ 的连接 $[X;L]$（作为正实例），以及句子序列 $x_1,\cdots,x_T$ 和生成器生成的标签序列 $y_1,\cdots,y_T$ 的连接 $[X;Y]$（作为负实例）. 当判别器训练到有了一定的判别能力，就可以为句子序列 $x_1,\cdots,x_T$ 和生成器生成的标签序列 $y_1,\cdots,y_T$ 的连接 $[X;Y]$ 打分了，并根据 $[X;Y]$ 得分与 $[X;L]$ 得分返回梯度给生成器，指导生成器提升训练. 具体过程见 3.5 小节.

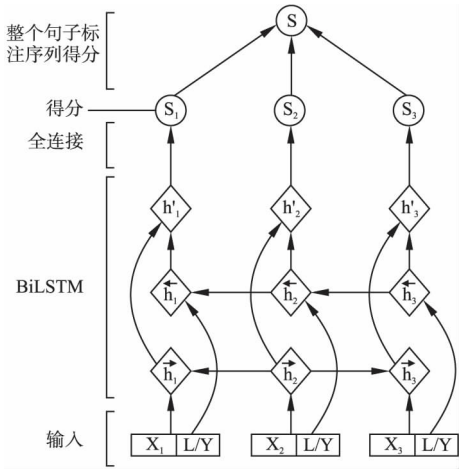

图 3 判别器的 BiLSTM 处理过程

Fig. 3 Bidirectional LSTM processing of discriminator

3.5 优化目标函数

本文提出的 CWGAN 模型使用 WGAN-GP 模型的梯度更新方式，即通过拉近真实样本分布和生成样本分布之间的 Wasserstein 距离（又叫 Earth-Mover，EM 距离）来优化目标函数. Wasserstein 距离的公式定义如下所示：

$$W(P_r,P_g)=\inf_{\gamma\sim\prod(P_r,P_g)}E_{(x,y)\sim\gamma}[\|x-y\|] \tag{4}$$

其中，P_r 代表真实样本的分布，P_g 代表生成器生成样本的分布，$\prod(P_r,P_g)$ 代表真实样本分布 P_r 和生成样本分布 P_g 组合起来的所有可能的联合分布的集合. 从每一个可能的联合分布 γ 中采样得到一个真实样本 x 和一个生成样本 y，即 $(x,y)\sim\gamma$，算出这对样本的距离 $\|x-y\|$，然后就可以计算该联合分布 γ 下样本对距离的期望 $E_{(x,y)\sim\gamma}[\|x-y\|]$. 在所有可能的联合分布中能够对这个期望值取到的下界（公式(5)等号右边），即定义为 Wasserstein 距离. 通过拉近真实样本和生成样本的 Wasserstein 距离来拉近两个样本的分布，其好处是在两个样本分布无重叠或重叠部分可忽略的情况下，Wasserstein 距离仍然可以提供有意义的梯度.

在对抗训练的过程中，判别器的作用是为生成器生成的数据（分布）打分，生成器则根据判别器给出的分数做出微小的调整，然后再将新生成的数据交给判别器打分. 所以当输入判别器的样本稍微改变，判别器不能给出与上次样本差距太大的分数，即需要限制判别器打分的变动幅度. 对于判别器的这种限制可以通过施加 Lipschitz 限制实现，如下：

$$\|f(x_1)-f(x_2)\|\leqslant K|x_1-x_2| \tag{5}$$

其中 K 是一个大于等于 0 的常数，x_1 和 x_2 是样本空间内的元素. 本文的 CWGAN 模型利用梯度惩罚来实现 Lipschitz 限制，即额外设置一个损失项：

$$[\|\nabla_x D(x)\|_p-K]^2 \tag{6}$$

为了使损失项的期望能够进行采样，只在生成样本集区域、真实样本集中区域以及夹在它们中间的区域进行采样. 加上以上损失项后判别器损失函数定义如下：

$$L(D)=-E_{x\sim P_r}[D(x)]+E_{x\sim P_g}[D(x)]+\lambda E_{x\sim P_{\hat{x}}}[\|\nabla_x D(x)\|_p-1]^2 \tag{7}$$

式中 K 设为 1，x 是判别器的输入样本，即连接序列 $[X;Y]$、$[X;L]$，P_r 代表真实样本 $[X;L]$ 的分布，P_g 代表生成器生成样本 $[X;Y]$ 的分布，$P_{\hat{x}}$ 代表在 x_r 和 x_g 的连线上随机插值采样 $\hat{x}$ 所满足的分布，$\hat{x}$ 可如下获得：

$$\hat{x}=\varepsilon x_r+(1-\varepsilon)x_g \tag{8}$$

判别器期望拉大两个分布之间的 Wasserstein 距离，生成器希望拉近两个分布之间的 Wasserstein 距离.

生成器的损失函数定义如下：

$$L(G)=E_{x\sim P_g}[D(x)]+cost \tag{9}$$

其中，$cost$ 代表真实样本与生成样本的交叉熵，计算公式如下：

$$cost=-[D((l')\log(D(y')))+(1-D(l'))\log(1-D(y'))] \tag{10}$$

其中，l' 代表句子序列和对应的标注标签序列的连接 $[X;L]$，y' 代表句子序列和生成器生成的标签序列的连接 $[X;Y]$.

本模型采用 Adam 优化算法来优化判别器和生成器的损失函数. 为了防止生成器训练时发生梯度爆炸，使用裁剪（clip）的方法限制每次更新后梯度的范围，一旦生成器的梯度超过了设定阈值就对其进行“裁剪”，使其保持在设定阈值范围内.

4 实验结果及分析

为了证明本文提出的 CWGAN 模型的优越性，本节设置了几组对比实验，通过比较 CWGAN 模型和 BiLSTM 模型在不同数据集上实现 NER 任务时的性能，以及不同设置下 CWGAN 模型在 NER 任务中的性能来说明 CWGAN 模型的效果. 另外需要说明的是，本文的 CWGAN 算法只与基础的 BiLSTM 模型进行了比较，因为上节提到的 BiLSTM + CRF 模型、CNN + BiLSTM 模型，以及 CNN + Attention + BiLSTM 模型都是在 BiLSTM 基础模型上增加新的模块从而改善了性能，本文算法同样可以在输入端和输出端增加相应的模块来改善性能，这里就不再一一进行比较.

4.1 数据集及评估标准

本节实验使用的数据集是 CoNLL-2002 中的西班牙文数据集和 CoNLL-2003 中的英文数据集.

CoNLL-2002 中 NER 任务数据集包含了西班牙文数据集和荷兰文数据集. 西班牙文数据集是由西班牙的 EFE 通讯社提供的新闻组成. 该数据集标记有四种不同的命名实体类型，分别为：人名（PERSON），地名（LOCATION），组织机构名（ORGANIZATION）以及其他命名实体（MISC），即不属于以上三种实体中的任何一种. 该数据集包含了标准的训练集，验

证集和测试集. 如表 1 所示.

表 1　CoNLL-2002 NER 任务西班牙文数据集规模表

Table 1　CoNLL-2002 NER Task Spanish dataset scale table

	人名	地名	组织机构名	其他	单词数量
训练集	8224	6804	12382	5385	273037
验证集	2081	1321	3066	1099	54837
测试集	1369	1409	2504	896	53049

CoNLL-2003 NER 任务数据集由路透社 RCV1 语料库的新闻专线组成. 它标有四种不同类别的命名实体类型: 人名(PERSON),地名(LOCATION),组织机构名(ORGANIZATION)以及其他命名实体(MISC). 该数据集包括标准的训练集,验证集和测试集. 如表 2 所示.

表 2　CoNLL-2003 NER 任务英文数据集规模表

Table 2　CoNLL-2003 NER Task English dataset scale table

	人名	地名	组织机构名	其他	单词数量
训练集	11135	8297	10027	4593	204568
验证集	3150	2094	2092	1268	51597
测试集	2777	1925	2496	918	46667

采用 NLP 任务中常用的评测指标 F-1 测度值对实验结果进行评价分析,F-1 测度值是对准确率和召回率的一种平均加权,它能够体现整体测试效果. 它的计算方法为:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

其中,P 代表准确率,P = 正确识别的命名实体个数/识别的命名实体总数 ×100%,R 代表召回率,R = 正确识别的命名实体个数/数据集中命名实体总数 ×100%.

4.2　实验设置

1) 预训练的词向量. 与随机初始化的词向量相比,使用预训练的词向量可以取得更好的效果.

本实验使用了两种预训练的词向量: Word2vec[5] 和 GloVe[6]. GloVe 与 Word2vec 都是基于词共现结构以无监督的方式学习单词的向量表示. 不同的是,GloVe 是对“词-词”矩阵进行分解从而得到词表示的方法,属于基于矩阵的分布表示,它相比 Word2vec 充分考虑了词的共现情况.

由于两者都具有比较优秀的准确性,并且这两种词向量是当前使用比较广泛的两种词向量,所以,此次实验采用了这两种方法生成的向量作为实验的输入. 本实验中使用 300 维的实值向量表示单词的词向量,word2vec 和 GloVe 训练词向量时使用的参数如表 3 所示.

表 3　词向量训练参数表

Table 3　Word vector training parameters table

词向量工具	词向量维度	窗口大小	学习率	采样阈值
Word2vec	300	3	0.01	1e-4
GloVe	300	3	0.01	—

2) 参数设置. 本文在训练时使用三折交叉验证法调整模型. 用网格搜索法来确定最优参数,并指定参数空间子集为: 窗口大小 $w \in \{1,2,3\cdots7\}$,过滤器数量 $n \in \{64,128,256,512\}$,生成器梯度裁剪阈值 $\in \{8,9,10,11,12\}$,随机梯度下降学习率 $\lambda \in \{0.1,0.01,0.001,0.0001\}$,使用 Adam 优化器更新参数. 本文实验使用的参数如表 4 所示.

表 4　实验参数表

Table 4　Table of experimental parameters

词嵌入维度	隐藏层神经元个数	窗口大小	批大小	丢弃率	学习率	梯度阈值
$d_w = 300$	$n = 256$	$w = 3$	$B = 16$	$p = 0.5$	$\lambda = 0.0001$	10

4.3　实验对比及分析

本文提出的 CWGAN 模型与 BiLSTM 模型[12] 的性能进行了对比. 表 5 是 BiLSTM 与 CWGAN 的基于西班牙文数据集的对比实验的结果. 表 6 是 BiLSTM、CWGAN 的基于英文数据集的对比实验结果. CWGAN 代表生成器和判别器都使用 BiLSTM. 由于西班牙用于预训练的数据不充足,导致实验效果不是很好,但是,实验的对比效果并没有因此受到影响.

表 5　基于 CoNLL-2002 的 CWGAN 模型效果表

Table 5　CWGAN Model base CoNLL-2002 effect table

模型	Test_a_F1	Test_b_F1	收敛所需的迭代次数
BiLSTM	42.89%	37.41%	38
CWGAN(BiLSTM-BiLSTM)	43.37%	38.61%	38

从表 5 中可以看出,在对于基于 CoNLL-2002 的实验上,无论是在验证集 Test_a 上还是测试集 Test_b 上,CWGAN 模型的 F1 值都比 BiLSTM 模型的 F1 值有所提高. 在验证集 Test_a 上,CWGAN 模型比 BiLSTM 模型提高 0.49%; 在测试集 Test_b 上,CWGAN 模型比 BiLSTM 模型提高 1.20%. 这说明 CWGAN 模型将生成对抗式网络用于命名实体识别任务是成功的,判别器能够指导生成器学习.

表 6　基于 CoNLL-2003 的 CWGAN 模型效果表

Table 6　CWGAN Model effect base CoNLL-2003 table

模型	Test_a_F1	Test_b_F1	收敛所需的迭代次数
BiLSTM	92.73%	88.01%	60
CWGAN(BiLSTM-BiLSTM)	93.02%	88.32%	40

从表 6 中可以看出,对于基于 CoNLL-2003 数据集的实验上,无论是在验证集 Test_a 上还是测试集 Test_b 上,CWGAN 模型的 F1 值比 BiLSTM 模型的 F1 值有所提高. 在验证集 Test_a 上,CWGAN 模型比 BiLSTM 模型提高 0.29%; 在测试集 Test_b 上,CWGAN 模型比 BiLSTM 模型提高0.21%. 这说明 CWGAN 模型将生成对抗式网络用于命名实体识别任务是成功的,判别器能够指导生成器学习.

从收敛迭代次数方面比较,CWGAN 模型的效果也优于

[5] https://en.wikipedia.org/wiki/Word2vec

[6] https://nlp.stanford.edu/projects/glove/

BiLSTM 模型,CWGAN 模型在迭代 40 次上下时达到收敛,而 BiLSTM 模型则需要迭代 60 次上下.

4.4 预训练词向量和 dropout 的影响

为了验证预训练的词向量以及 dropout 对于命名实体识别模型的影响,本小节做了以下对比实验. "dropout"是指在训练的时候,按一定的概率 p 来对权重层的参数进行随机采样.

表 7 基于 CoNLL-2003 数据集的不同设置下的 CWGAN 模型效果对比表

Table 7 CWGAN model effect comparison table under different Settings base CoNLL-2003 database

模型	Test_a_F1	Test_b_F1
CWGAN	84.17%	80.45%
CWGAN + dropout	85.81%	81.38%
CWGAN + pretrain(word2vec)	91.85%	86.99%
CWGAN + pretrain(GloVe)	73.60%	69.44%
CWGAN + pretrain(word2vec) + dropout	93.02%	88.32%
CWGAN + pretrain(GloVe) + dropout	74.62%	70.67%

表 7 显示了使用预训练的词向量和使用随机初始化的词向量的结果对比,表中"dropout"代表训练时设置丢弃率,"pretrain(word2vec) "代表使用 word2vec 工具预训练的词向量,"pretrain(GloVe) "代表使用 GloVe 工具预训练的词向量,CWGAN 和 CWGAN + dropout 代表使用随机初始化的词向量. 结果显示,使用预训练的 word2vec 词向量比使用随机初始化的词向量 F1 值提高了 6.99% 到 7.16%,说明与随机初始化的词向量相比,使用预训练的 word2vec 词向量可以获得更好的效果. 而使用 GloVe 词向量 F1 值反而比随机初始化的词向量效果更差了,因此建议预训练词向量时使用 word2vec 词向量. 表 7 中可以看出,在训练时使用"dropout"比不使用 F1 值提高了 1.17% 到 1.02%,说明在训练时使用"dropout"可以提高模型的性能,这是因为"dropout"在训练阶段可以阻止神经元的共适应.

5 结 论

本文提出了一个生成式对抗网络模型(CWGAN) 用于命名实体识别任务. 该网络模型借鉴 CGAN 以文本描述为条件的图像概率分布的思想,来完成命名实体识别以句子序列为条件获得标注序列概率分布的任务. 另外,该模型采用 WGAN-GP 中的梯度惩罚来保证梯度在后向传播的过程中保持平稳. 实验证明,本文提出的 CWGAN 模型在命名实体识别任务中是有效的,在对抗学习的过程中判别器可以指导生成器进一步提高自己的性能.

References:

[1] Chen Yu,Zheng De-quan,Zhao Tie-jun. Chinese relation extraction based on deep belief nets [J]. Journal of Software,2012,23(10) : 2572-2585.

[2] Grishman R. The NYU System for MUC-6 or where´s the syntax [C]//Proceedings of the 6th Conference on Message Understanding,Columbia,Maryland: Association for Computational Linguistics,1995: 167-175.

[3] Borthwick A,Grishman R. A maximum entropy approach to named entity recognition [D]. New York: New York University,Graduate School of Arts and Science,1999: 38-47.

[4] Lafferty J D,Mccallum A,Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Eighteenth International Conference on Machine Learning,Williams College,Williamstown,MA,USA: Morgan Kaufmann Publishers Inc,2001: 282-289.

[5] Zhou G D,Su J. Named entity recognition using an HMM-based chunk tagger [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,Pennsylvania,USA: Association for Computational Linguistics,2002: 473-480.

[6] McCallum A,Li W. Early results for named entity recognition with conditional random fields,feature induction and web-enhanced lexicons [C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4,Stroudsburg,USA: Association for Computational Linguistics,2003: 188-191.

[7] Putthividhya D P,Hu J. Bootstrapped named entity recognition for product attribute extraction [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,UK: Association for Computational Linguistics,2011: 1557-1567.

[8] Etzioni O,Cafarella M,Downey D,et al. Unsupervised named-entity extraction from the web: an experimental study [J]. Artificial Intelligence,2005,165(1) : 91-134.

[9] Li Li-shuang,He Hong-lei,Liu Shan-shan,et al. Research of word representions on biomedical named entity recongnition [J]. Journal of Chinese Computer Systems,2016,37(2) : 302-307.

[10] Banko M,Cafarella M J,Soderland S,et al. Open information extraction from the web [C]//International Joint Conference on Artificial Intelligence (IJCAI) ,Banff,Canada,2007,7: 2670-2676.

[11] Etzioni O,Fader A,Christensen J,et al. Open information extraction: the second generation [C]//International Joint Conference on Artificial Intelligence (IJCAI) ,Barcelona,Catalonia,Spain,2011,11: 3-10.

[12] Athavale V,Bharadwaj S,Pamecha M,et al. Towards deep learning in hindi NER: an approach to tackle the labelled data scarcity [C]//13th International Conference on Natural Language Processing. IIT(BHU) ,Varanasi,2016: 154-162.

[13] Huang Z,Xu W,Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. Computer Science,2015,42(5) : 45-54.

[14] Lample G,Ballesteros M,Subramanian S,et al. Neural architectures for named entity recognition [C]// The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) ,2016: 260-270.

[15] Chiu J P C,Nichols E. Named entity recognition with bidirectional LSTM-CNNs [J]. Transactions of the Association for Computational Linguistics,2016,4(17) : 357-370.

[16] Rei M,Crichton G,Pyysalo S. Attending to characters in neural sequence labeling models [C]//Proceedings of COLING 2016,the 26th International Conference on Computational Linguistics: Technical Papers,Osaka,Japan,2016: 309-318.

[17] Goodfellow I J,Pouget-Abadie J,Mirza M,et al. Generative adversarial nets [C]//International Conference on Neural Information Processing Systems,MIT Press,2014: 2672-2680.

[18] Yu L,Zhang W,Wang J,et al. SeqGAN: sequence generative adversarial nets with policy gradient [C]//11 Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence,AAAI Press,2017: 2852-2858.

[19] Arjovsky M,Chintala S,Bottou L. Wasserstein GAN [J]. arXiv Preprint arXiv: 1701.07875,2017.

[20] Mirza M,Osindero S. Conditional generative adversarial nets [J]. arXiv preprint arXiv: 1411.1784,2014: 2672-2680.

[21] Gulrajani I,Ahmed F,Arjovsky M,et al. Improved training of wasserstein gans [C]//Advances in Neural Information Processing Systems,Long Beach Convention Center,Long Beach,2017: 5769-5779.

附中文参考文献:

[1] 陈 宇,郑德权,赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取 [J]. 软件学报,2012,23(10) : 2572-2585.

[9] 李丽双,何红磊,刘珊珊,等. 基于词表示方法的生物医学命名实体识别 [J]. 小型微型计算机系统,2016,37(2) : 302-307.