融合自编码器和对抗训练的中文新词发现方法

潘韦[1,2]，刘天元[1,2]，孙宇清[1,2]，龚斌[1,2]，张永满[3]，杨萍[4]
1.山东大学，软件学院，济南，250101
2.山东大学，教育部数字媒体技术工程研究中心，济南，250101
3.曙光云计算集团有限公司，北京，100193
4.航天科工智慧产业发展有限公司，北京，100094

摘要

新词的不断涌现是语言的自然规律，如在专业领域中新概念和实体名称代表了专业领域中某些共同特征集合的抽象概括，经常作为关键词在句子中承担一定的角色。新词发现问题直接影响中文分词结果和后继文本语义理解任务的性能，是自然语言处理研究领域的重要任务。本文提出了融合自编码器和对抗训练的中文新词发现模型，采用字符级别的自编码器和无监督自学习的方式进行预训练，可以有效提取语义信息，不受分词结果影响，适用于不同领域的文本；同时为了引入通用语言学知识，添加了先验句法分析结果，借助领域共享编码器融合语义和语法信息，以提升划分歧义词的准确性；采用对抗训练机制，以提取领域无关特征，减少对于人工标注语料的依赖。实验选择六个不同的专业领域数据集评估新词发现任务，结果显示本文模型优于其他现有方法；结合模型析构实验，详细验证了各个模块的有效性。同时通过选择不同类型的源域数据和不同数量的目标域数据进行对比实验，验证了模型的鲁棒性。最后以可视化的方式对比了自编码器和共享编码器对不同领域数据的编码结果，显示了对抗训练方法能够有效地提取两者之间的相关性和差异性信息。

关键词： 中文分词 ； 新词发现 ； 自编码器 ； 对抗框架

Finding Chinese New Word By Combining Self-encoder and Adversarial Training

Wei Pan[1,2],Tianyuan Liu[1,2],Yuqing Sun[1,2], Bin Gong[1,2],Yongman Zhang[3],Ping Yang[4]
1.Shandong University,School of Software,Jinan,250101
2.Shandong University,ERC of Digital Media Technology, MoE,Jinan,250101
3.Dawning Cloud Computing Group Co.,Ltd,Beijing,100193
4.CASIC Intelligence Industry Development Co.,Ltd,Beijing,100094

Abstract

The continuous emergence of new words is a natural law of language. Especially in professional fields, the new concepts and entity names in sentences are abstraction of professional features and often take functions as keywords. The problem of new word discovery not only directly affects the results of Chinese word segmentation, but also seriously affects the semantic understanding of text and the performance of subsequent tasks. It is an important task in the field of natural language processing. This paper proposes a Chinese new word discovery model that combines self-encoder and adversarial training. The character-level self-encoder can be pre-trained through unsupervised self-learning, which can effectively extract semantic information without

基金项目： 山东省自然科学基金(ZR2018ZB0420);国家重点研发计划(2018YFC0831401,2018YFC0831406);国家自然科学基金 (91646119);山东省重点研发计划(2019JZZY010107)。

being affected by word segmentation results, so it is suitable for different fields. In order to introduce linguistic knowledge in the general field, a prior syntactic analysis result is added. Semantic and grammatical information is merged with the help of the domain sharing encoder to improve the accuracy of divergent words. The adversarial training is used to extract domain-independent features and reduce the dependence on manual annotation corpus. Experiments are conducted on six different professional domain data sets to evaluate the new word discovery task. The results show that the performances of our model are better than other methods. Through the model ablation experiments, the effectiveness of each module is verified in detail. The robustness of the model are evaluated against different types of source domain data and different amounts of target domain data. Finally we visually compare the encoding results of the self-encoder and the shared encoder for different fields of data, which shows the adversarial training method can effectively extract the correlation and difference information between the two domains.

Keywords: Chinese word segmentation , New word discovery , Self-encoder , Adversarial framework

1 引言

新词的不断涌现是语言的自然规律，随着科技进步和社会经济发展，描述新事物的词汇越来越多，这些词汇包括名词、动词、形容词等多种形式，如指代人名、地名、组织机构名、产品名称、书籍或电影名称的专有名词。尤其是在专业领域，不断出现新概念和实体名，描述了领域中涌现的新事物或共同特征集合的抽象表述，例如医学领域的专业词汇“吲哚美辛”。这些领域新词相比于其他通用单词，经常作为关键词在句子中承担一定的角色，对于需要分词的中文文本，新词带来的分词错误将直接影响文本语义的理解，影响到后续文本处理任务性能。

目前新词发现方法主要分为两类，一是依据词库进行统计进行分词的方法，但是此类方法因词库更新不及时，无法应对专业领域新词大量增长的情况，并且基于规则的统计方法不适合跨领域应用，不适用于具有不同特点的专业领域；二是基于神经网络的分词方法，需要大量人工标注语料库，难以用于发展迅猛和差异大的专业领域，难以辨别歧义词，新词标注准确率不高。

针对上述问题，本文提出融合自编码器和对抗训练的中文新词发现模型框架。在模型中引入预训练自编码器以提取句子级别的语义信息；添加先验句法知识提升歧义词的划分准确性；使用共享编码器融合语义和语法信息以及跨领域特征；最后采用对抗训练的机制，提取领域无关的特征，进行新词发现。本文在六个不同的专业领域数据集上评估新词发现任务，其结果优于其他的对比方法；进行了模型析构实验，验证模型中各个模块的有效性。通过在不同类型的源域数据和不同数量的目标域数据进行对比实验，以验证模型鲁棒性。最后可视化显示了不同领域之间特征的相关性和差异性。

2 相关工作

2.1 分词方法

目前中文分词方法主要可以分为两类:统计方法和神经方法。基于规则和统计的方法是依据语言学知识整理的模板，通过程序匹配进行新词检测，使用的统计特征包括字符共现频率，左右邻接字的互信息，邻接熵等。例如，隐马尔可夫模型(HMM) (Schulz, 1992)，通过可观测状态集合和这些状态与隐含状态之间的概率关系，计算已知序列的最大转移概率，采用维特比算法动态规划的思想，给出分词结果。条件随机场模型(CRF) (Lafferty et al., 2001)是最大熵和马尔可夫模型的结合，不仅考虑了字符共现频率等统计特征，还融合了上下文信息，通过全局最优化进行分词。陈飞等人 (陈飞，刘奕群，魏超, 2013)对新词边界特征提出一些划分规律，并采用CRF方法拟合特征进行新词划分，但这种统计方法存在计算复杂度高、训练代价大等问题。

基于神经网络的方法常采用端到端序列标记神经架构，在标记样本集合上进行监督学习。例如Chen等人 (Chen and Qiu, 2015)首次使用长短期记忆神经网络(LSTM)网络来解决中文分词中的长期依赖关系问题。Gong等人 (Gong and Chen, 2019)提出了一种用于中文分词的灵活的多准则学习，它由几个长短期记忆神经网络和一个自动在这些长短期记忆神经网络之间切换路由的切换器组成，提供了更灵活的解决方案。Yang等人 (Yang et al., 2018)使用子词嵌入集成字符序列上的Lattice LSTM网络，来研究用于中文分词的子词信息。Zhao等人 (Zhao et al., 2018)提出了一种合并未标记和部分标记的数据的模型以利用未标记的数据，作者使用门机制将双向LSTM分段模型与两个字符级语言模型结合在一起，同时修改了RNN的原始交叉熵损失函数。Cai等人 (Cai et al., 2017)提出了一种具有平衡词和字符嵌入输入的贪婪神经网络分词器，来减轻当前神经模型的训练和工作过程在计算上低效的缺点，进行更快更准确的分词。Ye等人 (Ye and Zhang, 2019)提出了一种基于词的半监督方法，采用子采样和负采样方法来导出针对分词优化的词嵌入以改善跨域分词的效果。这些神经网络方法可以很好地从语料中学习数据特征进行分词，但需要大量的标记数据来进行模型训练。

为了解决上述问题，部分工作在神经网络的基础上引入外部字典等其他知识辅助分词，例如Liu等人 (Liu et al., 2018b)提出了两种方法来利用字典信息，第一种是基于伪标记数据生成的方法，而第二种方法则是基于多任务学习。该方法在训练语料缺乏的情况下取得了较好的效果。Liu等人 (Liu et al., 2018a)认为传统的新词发现方法只能找到分词工具词典中不存在的词，但是这些单词不一定会影响单词分割的结果。因此作者提出基于分割结果构建候选新单词的集合，以这种方式发现所有被分词工具错分的新单词，并添加到分词工具中。Zhang等人 (Zhang et al., 2018)提出了两种扩展双向长短期记忆神经网络的方法将字典合并到神经网络中以解决中文分词任务的问题。该方法具有处理稀有词和训练数据不同造成结果偏差的问题，利用先验知识进行领域转移来提升效果。Wang等人 (Wang et al., 2018)分析多种字符嵌入方式，包括拼音和五笔输入法，提出了一种新颖的共享Bi-LSTM-CRF模型，可以通过在训练过程中共享LSTM网络来有效地融合语言特征包括字符语义和语音含义。

2.2 对抗训练

对抗训练是一种通过引入噪声数据进行模型训练的方式 (Goodfellow et al., 2014)，以此来增强神经网络模型的鲁棒性和提升模型的泛化能力。对抗训练方式首先应用于计算机视觉领域，并取得了良好效果。近年来，对抗训练的方法也应用到自然语言处理领域，例如Cao等人 (Cao et al., 2018)提出一个新颖的对抗训练模型，将命名实体识别和分词两个任务进行联合学习。作者充分利用两个任务中共享边界的信息，显式捕获两个字符间远程依赖关系，使每个字符都可以在预测字符类型时提供有用的信息。Ding等人 (Ding et al., 2020)重新考虑了汉语单词的本质，并设计了自动远距离标记机制，不需要任何来自目标域的监督或预定义词典，并使用句子级别的对抗训练来进行源域信息的降噪。Chen等人 (Chen and Shi, 2017)针对多标准分词问题进行研究，利用不同标准的共享知识进行多准则对抗学习分词，效果相对其他单准则学习方法有所提高。

3 融合自编码器和对抗训练的中文新词发现模型

针对专业领域新词发现问题，本文提出了融合自编码器和对抗训练的中文新词发现模型，包括三部分，如图1所示。第一部分是基于文本重构的自编码器，采用无监督预训练的方式提取句子级别的语义信息；第二部分则是添加先验句法知识，与字符向量融合，提升歧义词划分的准确性；第三部分则引入对抗训练的机制，利用各领域间存在共性的特征，解决专业领域标注数据较少问题；最后采用CRF结构对字符序列进行标注，进行有效的新词发现。

3.1 预训练文本重构自编码器

本文采用字符级别的自编码器提取文本语义信息，如图1中的上半部分所示。自编码器包括编码器和解码器两部分，编码器将输入序列编码为隐式空间向量$\boldsymbol{H}$，编码器和解码器可以有很多选择如LSTM，BiLSTM，CNN等。解码器采用自学习方式将该向量$\boldsymbol{H}$进行重构为原文本。通过对输入数据重构，可以使$\boldsymbol{H}$包含句子的语义信息，是后续分词过程中语义融入的基础。

自编码器首先在海量语料库进行训练，本模型采用维基百科数据语料无监督地进行网络预训练。这种使用预训练的方法，可以为模型提供了一个良好的初始化参数，在语义信息编码任

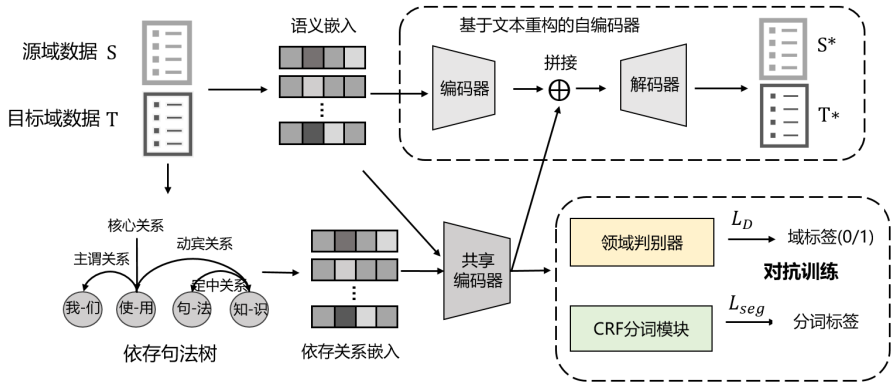


图 1: 融合自编码器和对抗训练的新词发现模型

务上可以有更好的泛化性能、并加速收敛训练速度；同时可以将在通用语料库中学习到的语言学特征融入，获得当前输入句子的完整信息。此外，预训练好的模型，可以有效避免在小数据集上发生过拟合现象，提供一种正则化的功能。

本文模型所接受的输入为字符序列，并选择双向循环神经网络BiLSTM作为编码器和解码器，BiLSTM是由前向及后向LSTM循环神经网络联合构成，前向LSTM是以正向的顺序对序列的每个字符进行编码，后向LSTM则相反之，最后将每一个字符对应位置的前向LSTM和后向LSTM的输出向量进行拼接。这种方式可以捕获目标字符上下文的语义信息，解决字符之间的长距离依赖问题。输入句子$s = c_1, c_2, \ldots, c_n$，首先通过嵌入层，将每个字符转化为字符向量$\boldsymbol{e}_i$，然后通过编码器得到双向拼接后的语义特征向量$\boldsymbol{h}_i$，公式如1-3。

$$\overrightarrow{\boldsymbol{h}}_i = LSTM_f(\boldsymbol{e}_i) \tag{1}$$

$$\overleftarrow{\boldsymbol{h}}_i = LSTM_b(\boldsymbol{e}_i) \tag{2}$$

$$\boldsymbol{h}_i = \overrightarrow{\boldsymbol{h}}_i \oplus \overleftarrow{\boldsymbol{h}}_i \tag{3}$$

在获得相应的语义特征向量后，将其输入到解码器中。解码器的作用是将特征向量$\boldsymbol{h}_i$转换为相应输入的字符，公式如4-5。

$$\boldsymbol{\nu}_i = BiLSTM_{self}(\boldsymbol{h}_i) \tag{4}$$

$$\hat{\boldsymbol{\mu}} = softmax(\boldsymbol{\nu}_i) \tag{5}$$

其中$\hat{\boldsymbol{\mu}} \in R^r$，$r$为字典大小。本文自训练结果准确率达到98.6%，准确率计算方法为字符正确个数除以样本字符总个数。

3.2 融入先验依存句法知识

句法知识不同于单词的在应用方面的灵活性，具有较为稳定的结构关系，可以直接在不同领域之间进行迁移并提供有用的信息，如Tian (Tian and Song, 2020)等人使用了双向注意机制来结合上下文特征和输入字符的相应句法知识，并在文中证明了模型可以从先验句法知识中获益。

本文使用百度自然语言处理部的DDParser (Zhang and Wang, 2020)依存句法分析工具获得句子的依存关系，如图2所示的“他向我们说明文中的细节”完整的依存句法结构图。在图中，没有将“说明文”作为一个单词和其他短语组合为不同的关系。从这个例子里可以看出使用依存句法知识可以减少分词过程中的歧义词问题。

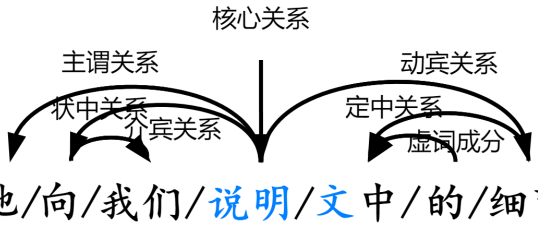


图 2: 中文分词例句，蓝色标注内容为歧义词部分，弧线表示依存句法关系

3.3 领域对抗训练

在模型结构中，源域和目标域语义信息获取的过程是独立的，但两个领域仍存在可共享的内容。受对抗训练相关工作的启发，通过将输入源域和目标域的混合数据融入到共享层中，使用对抗架构以确保共享层可以生成与领域无关的特征向量，提取与领域无关的信息，增强模型的鲁棒性和泛化性。

共享编码器的结构使用双向循环神经网络BiLSTM，编码器的输入是源域句子$s_{src} = c_1^s, c_2^s, \ldots, c_n^s$和目标域句子$s_{tgt} = c_1^t, c_2^t, \ldots, c_m^t$的混合数据，以及每个字符之间的依存关系$s_{dep} = dep_1, dep_2, \ldots, dep_p$，其中$c_i^s$表示长度为$n$的源域句子$s_{src}$第$i$个字符，$c_j^t$表示长度为$m$的目标域句子$s_{tgt}$第$j$个字符。这些字符序列也被输入到嵌入层，转化为字符向量，如源域字符向量序列$\boldsymbol{E}_{src} = \boldsymbol{e}_1^s, \boldsymbol{e}_2^s, \ldots, \boldsymbol{e}_n^s$，目标域字符向量序列$\boldsymbol{E}_{tgt} = \boldsymbol{e}_1^t, \boldsymbol{e}_2^t, \ldots, \boldsymbol{e}_m^t$，依存关系向量序列$\boldsymbol{E}_{dep} = \boldsymbol{e}_1^{dep}, \boldsymbol{e}_2^{dep}, \ldots, \boldsymbol{e}_p^{dep}$，其中$\boldsymbol{e}_i^s$表示源域句子第i个字符的字符向量，$\boldsymbol{e}_j^t$表示目标域句子第j个字符的字符向量，$\boldsymbol{e}_k^{dep}$表示句子第k个字符的依存关系向量。最后将字符向量与依存关系向量拼接，输入到共享编码器中。

$$\boldsymbol{e}_i^* = \boldsymbol{e}_i^{s/t} \oplus \boldsymbol{e}_i^{dep} \tag{6}$$

$$\boldsymbol{h}^*{}_i = BiLSTM(\boldsymbol{e}_i^*) \tag{7}$$

根据上述公式6-7，得到源域和目标域的共享特征向量$\boldsymbol{H}_{src}^* = \boldsymbol{h}_1^{s*}, \boldsymbol{h}_2^{s*}, \ldots, \boldsymbol{h}_n^{s*}$和$\boldsymbol{H}_{tgt}^* = \boldsymbol{h}_1^{t*}, \boldsymbol{h}_2^{t*}, \ldots, \boldsymbol{h}_m^{t*}$，其中$\boldsymbol{h}_i^{s*}$表示源域句子第$i$个字符的共享特征向量，$\boldsymbol{h}_j^{t*}$表示目标域句子第$j$个字符的特征向量，$\boldsymbol{H}_{src}^* \in R^{n \times 2d}, \boldsymbol{H}_{tgt}^* \in R^{m \times 2d}, i \in [0, n], j \in [0, m]$，$d$表示共享编码器隐藏单元的数量。

对抗任务中定义的鉴别器在于区分每个句子所属的领域，并且已经进行充分的预训练，最终目的是希望共享编码器可以混淆鉴别器，提取共享特征信息。考虑本文模型的数据量不是很大，因此没有使用深度网络作为鉴别器，直接将共享编码层的输出接入最大池化层，去除冗余信息，进行特征压缩，以此来加快速度。最后将池化后的特征向量进行$Sigmoid$二分类，判断领域归属。计算公式8-9，其中$\boldsymbol{H}_{src/tgt}^*$表示共享编码层生成的隐式特征向量，$\boldsymbol{W}_d$和$\boldsymbol{b}_d$表示可训练的参数，$\theta_d$ 表示鉴别器参数。

$$\boldsymbol{g} = Maxpooling(\boldsymbol{H}_{src/tgt}^*) \tag{8}$$

$$D(\boldsymbol{g}; \boldsymbol{\theta}_d) = Sigmoid(\boldsymbol{W}_d \boldsymbol{g} + \boldsymbol{b}_d) \tag{9}$$

3.4 基于条件随机场的分词标记

本文在共享编码层后，使用条件随机场（CRF）输出分词结果。因为序列标注任务追求的不是单个字符预测的准确性，而更关注输入序列整体的合理性。因此，引入CRF概率图模型来搜索最符合条件的序列路径，考虑序列字符之间的关系，得到输出中的最优解。

由图1可知，该层的输入是共享编码层向量$\boldsymbol{H}^*$,其中$\boldsymbol{H}^* = \boldsymbol{h}_1^*, \boldsymbol{h}_2^*, \ldots, \boldsymbol{h}_{m/n}^*$，$\boldsymbol{h}_i^*$表示输入序列中第$i$个字符的共享特征向量。对于标签序列$y$的预测输出，使用$Softmax$激活函数选择最大

概率选项，其中Y表示输入序列X的所有可能的字符标签，W, b为可训练的参数，计算方式如公式10-11所示。

$$P(y|X) = \frac{\exp\left\{\sum_i \left\{W^{y_i} h_i + b^{(y_{i-1}, y_i)}\right\}\right\}}{\sum_{y' \in Y} \exp\left\{\sum_i \left\{W^{y'_i} h_i + b^{(y'_{i-1}, y'_i)}\right\}\right\}} \tag{10}$$

$$\hat{\boldsymbol{y}} = \underset{y \in Y}{agrmax} P\left(y|X\right) \tag{11}$$

3.5 模型训练

3.5.1 面向专业文本的自编码器微调训练

针对新领域应用，首先需要使用少量该领域标注语料进行微调文本重构自编码器。根据公式4-5，将预训练自编码器中源域和目标域的特征向量$\boldsymbol{H}_{src}$和$\boldsymbol{H}_{tgt}$与共享编码层对应领域的特征向量拼接$\boldsymbol{H}_s = [\boldsymbol{H}_{src} \bigoplus \boldsymbol{H}^*_{src}]$，$\boldsymbol{H}_t = [\boldsymbol{H}_{tgt} \bigoplus \boldsymbol{H}^*_{tgt}]$，其中$\boldsymbol{H}_s = \boldsymbol{h}_{s1}, \boldsymbol{h}_{s2}, \ldots, \boldsymbol{h}_{sn}$，$\boldsymbol{H}_t = \boldsymbol{h}_{t1}, \boldsymbol{h}_{t2}, \ldots, \boldsymbol{h}_{tm}$，这样在更新自编码器时，也会更新$\boldsymbol{H}^*_{src}$和$\boldsymbol{H}^*_{tgt}$，从而将语义信息融入到新词发现任务中，自编码器的形式更新为公式12-13，其中$\hat{\boldsymbol{\mu}}_{si/vi} \in R^r$，$r$为字典大小，$\boldsymbol{\theta}_{pre}$为预训练自编码器参数。

$$\boldsymbol{\nu}_{si/vi} = BiLSTM_{self}(\boldsymbol{h}_{si/vi}; \boldsymbol{\theta}_{pre}) \tag{12}$$

$$\hat{\boldsymbol{\mu}}_{si/vi} = softmax(\boldsymbol{\nu}_{si/vi}) \tag{13}$$

3.5.2 联合损失函数

本模型将自编码器，分词任务，对抗训练进行联合学习，模型的损失函数由这三部分共同组成。针对自编码器，要求输出的数据和输入保持一致，每个字符位置为最大概率类别，通常使用交叉熵作为自编码器的损失函数。

$$R^s = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}^s{}_i log(\hat{\boldsymbol{\mu}}^s_i) \tag{14}$$

$$R^t = -\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\mu}^t{}_j log(\hat{\boldsymbol{\mu}}^t_j) \tag{15}$$

其中$\hat{\boldsymbol{u}}^s_i \in R^r$表示模型计算得到的源域句子第i个字符在字典类别上的概率分布向量，$\boldsymbol{u}^s_i \in R^r$表示该字符类别的真实标签，r表示字典大小，n表示源域句子的长度。$\hat{\boldsymbol{u}}^t_i \in R^r$表示模型计算得到的目标域句子第j个字符在字典类别上的概率分布向量，$\boldsymbol{u}^t_i \in R^r$表示该字符类别的真实标签，$m$表示目标域句子的长度。

对于分词部分，则采用使用一阶维特比算法来计算标签序列的最优分数，字符标注的句子级别对数似然损失函数定义如公式16，其中G表示所有训练数据，包括源域和目标域，λ为正则化权重，$\boldsymbol{\Theta}$代表模型参数集合。

$$L_{seg} = -\sum_{G} log(\hat{\boldsymbol{y}}) + \lambda \|\boldsymbol{\theta}\|^2 \tag{16}$$

对于对抗训练部分，目的在于混淆鉴别器，让其无法区分共享层的特征向量是来着源域还是目标域，损失函数如公式17。其中D_s, D_t分别表示源域和目标域的全部数据，并且$d_i \in 0, 1$，在此定义标签0表示数来源于源域，标签1表示数据来源于目标域。$\hat{d}_i$表示sigmoid函数预测出来的概率。

$$L_D = \sum_{i=0}^{D_s+D_t} d_i log(\hat{d}_i) + (1 - d_i) log(1 - \hat{d}_i) \tag{17}$$

结合上述各个部分，可以得出最终损失函数如公式18，其中α表示自编码器损失权重，β表示分词部分损失权重，γ表示对抗训练部分损失权重。

$$L = \alpha(R^s + R^t) + \beta(L_{seg}) + \gamma L_D \tag{18}$$

4 实验设计与分析

4.1 数据集与评价指标

不同领域的文本具有很大差异，因此在实验中选择不同专业领域的数据集进行实验 (Qiu and Zhang, 2015) (Ye and Zhang, 2019)，包括新闻，文学，医学和专利领域。这些均是内容变化较快的领域，相较于其他数据集，可以包含更多的新生词汇，从而反映模型的效果。本文数据集的统计量如表1所示。在数据集中，PKU(北京大学)数据集，主要包括新闻领域内容，由于该文本内容较为正式，且新词不如其他领域数据多，所以将其作为源域数据集使用，帮助提取有效信息。DL(斗罗大陆), FR(凡人修仙传), ZX(诛仙)是小说数据集，其包含大量的专有新词，如人物名字，技能名称等。DM (皮肤学)为皮肤科学专业领域数据集，PT(专利)为专利领域数据集，其中都包含大量专业词汇，适用于本文的目标任务。

特别说明的是，为了评估新词发现任务，对目标域测试集的词汇表进行了统计，从中去除结巴字典中的单词，之后剩余的词汇为专业领域新词，新词个数如表1最后一列所示。为了后续实验的公正性，在进行新词发现任务时，模型的训练数据去除包含新词的语句。对于通用分词任务(G)和新词发现任务(N)均使用准确率作为评价指标，用于衡量分词模型的准确程度。通用分词任务的准确率表示句子中正确划分的单词个数N_{pw}与句子标注单词个数N_{sw}的比值的期望，即$Acc_G = E(N_{pw}/N_{sw})$。对于新词发现任务，由于不是每个句子都有新词，因此在分词任务完成后，统计测试集中正确识别出的领域新词个数N_{nw}与测试集新词总个数N_{tw}的比值，即$Acc_N = N_{nw}/N_{tw}$。

本实验所采用的字符向量以及依存关系向量均来自于Li等人 (Li and et al, 2018)共享的预训练向量集。本文模型运行在NVIDIA Geforce RTX 2080Ti上，使用Tensorflow框架并采用Adam优化器进行优化。

类型	数据集	领域	句子统计		单词统计		
			训练集	测试集	训练集	测试集	新词个数（测试集）
源域	PKU	新闻	47	-	52	-	-
源域/目标域	FR	文学	148	1	86	3.5	0.8
目标域	DL	文学	40	1	46	4.7	0.9
目标域	ZX	文学	59	1	39	3.8	0.9
目标域	DM	医学	32	1	33	3.2	1
目标域	PT	专利	17	0.7	31	4.3	1.1

表 1: 不同领域数据集统计(单位：千/K)

4.2 对比方法

- **MCCWS：** (Qiu and Pei, 2019)该模型以Transformer框架为基础，提出自注意力结构加CRF组合方式，模型输入除了字符向量和位置向量外，额外增加双字向量，增强字符之间的邻接关系。

- **CWS-DICT：** (Zhang et al., 2018)该模型引入外部字典，使用字符向量和特征向量作为输入，该特征向量表示n-gram窗口下是否为单词的one-hot表示。作者使用两个并行的 Bi-LSTM 用来提取上下文信息和潜在的字边界信息，将两个隐式向量拼接，输入到CRF层分词。

- **ASCWS：** (Chen and Shi, 2017)该模型在BiLSTM-CRF基础上引入对抗训练方式，融合多个标准分词方法，挖掘它们的共同特征，提升分词效果。

- **BiLSTM-CRF：** (Chen and Qiu, 2015)该模型使用BiLSTM对输入的字符序列编码，通过CRF层进行序列标注。该方法是神经网络中文分词中使用端到端序列标记架构的经典方法。

4.3 模型性能比较

在该任务中，使用新闻领域数据集PKU作为源域数据集，选择文学，皮肤学，专利领域作为目标域数据集进行训练，从表1中可以看出，目标域数据集的数量远小于源域的数据量。本实验在五个不同领域的数据集上进行通用分词和新词发现任务，与其他代表性的方法进行对比。

任务类型	方法名称	数据集				
		DL	FR	ZX	DM	PT
通用分词任务 Acc_G	BiLSTM-CRF	89.1	83.4	84.3	76.1	80.0
	MCCWS	91.7	84.7	85.4	77.4	83.4
	CWS-DICT	90.5	84.5	87.1	**78.2**	85.6
	ASCWS	93.0	85.2	88.7	77.6	85.0
	Our Model	**93.5**	**85.6**	**89.9**	78.0	**87.6**
新词发现任务 Acc_N	BiLSTM-CRF	62.7	60.8	63.2	52.5	50.7
	MCCWS	61.1	62.0	64.6	54.3	53.4
	CWS-DICT	61.2	63.1	66.1	55.4	52.5
	ASCWS	65.4	64.6	65.7	55.9	55.3
	Our Model	**66.3**	**65.2**	**67.9**	**57.4**	**57.6**

表 2: 不同领域的分词任务和新词发现任务的对比分析

结果显示表2通用分词任务中，本文模型在PT数据集上相比于BiLSTM-CRF经典方法提升了7.6%，这表明相较于传统的神经网络分词方法具有很大的提升。CWS-DICT方法引入专业词典，在DM数据集上取得了最高分78.2%。但相比于本方法，CWS-DICT方法需要引入外部字典来判断n-gram分割的字符是否为单词，这种做法虽然提高了模型的效果，但是极大的约束了模型的迁移性。当在新领域并没有很全面的字典时，会严重降低模型的效果；而本模型引入源域和目标域双域进行学习，充分利用源域数据特征，可以更好地进行跨域分词。ASCWS方法是除了本模型，表现最好的基线模型。相比于ASCWS方法，本文添加了预训练自编码器和句法知识，可以在分词时利用提取到的句子结构及语义信息，提高分词效果。

在新词发现任务中，本模型在所有领域数据集上都取得了领先的结果。BiLSTM-CRF，MCCWS，CWS-DIC这些方法属于单领域方法，即只使用同一个领域的数据进行训练和测试。本模型采用对抗领域的学习方法，可以在足够多的源域数据上进行与领域无关特征抽取，辅助目标域新词发现,因此优于这些单领域方法。相比于BiLSTM-CRF，MCCWS，CWS-DIC模型，本文的方法准确率分别最高提升了6.9%，5.2%和5.1%。

4.4 模型析构分析

为了验证模型各个模块的有效性，按照图1的结构划分，对自编码器结构，句法知识融入模，对抗任务以及预训练效果四个模块进行有效拆分，并在文学，医学和专利三个领域分别进行通用分词和新词发现任务。

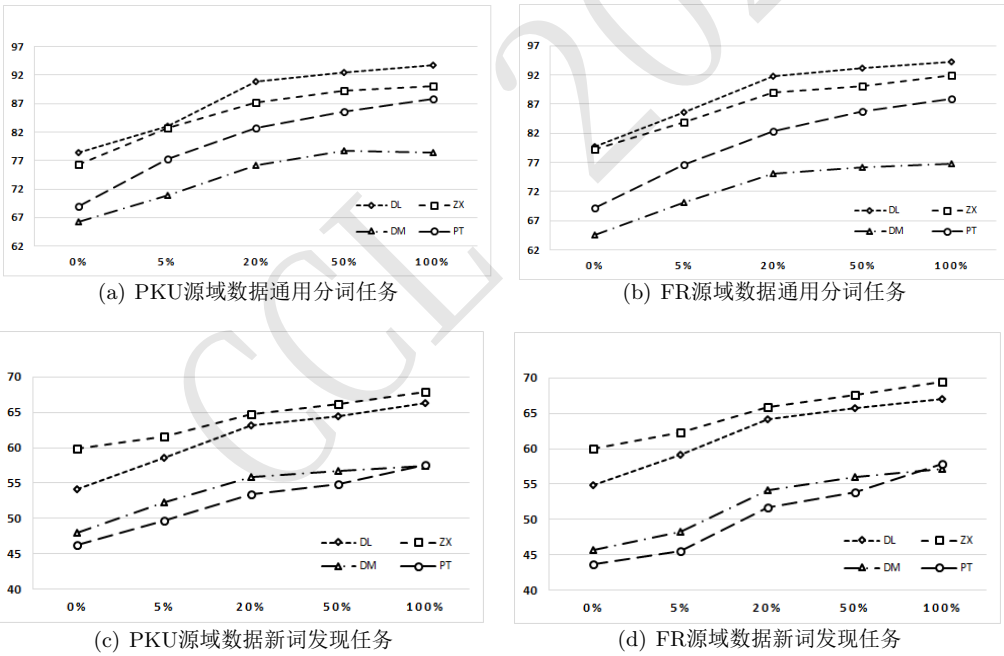
如表3所示，模型中的各个结构都起到了正向作用。其中对抗训练模块作用最大，在添加对抗训练任务后，在三个领域数据集上通用分词任务指标分别提升了2.2%，3.5%和3.6%，新词发现任务指标分别提升了3.5%，4.4%和3.1%。这是由于对抗模块可以对共享编码器进行训练，获

模型析构	DL		DM		PT	
	Acc_G	Acc_N	Acc_G	Acc_N	Acc_G	Acc_N
(无)自编码器	92.0	63.2	75.9	54.6	84.2	55.1
(无)句法知识	93.1	65.6	76.3	56.3	86.6	56.8
(无)对抗训练	91.3	62.8	74.5	53.0	84.0	54.5
(无)预训练	92.3	64.4	76.0	55.2	85.8	55.9
本文完整模型	**93.5**	**66.3**	**78.0**	**57.4**	**87.6**	**57.6**

表 3: 模型析构实验结果分析

取领域无关特征，帮助目标专业领域分词，提升新词识别准确度。自编码器的影响排在第二位，在添加自编码器模块后，在各个领域上也出现了评价指标上升的情况，这也说明了自编码器可以帮助融入句子的语义信息，并且通过预训练自编码器可以进一步提升模型效果，证明了自训练可以在语义信息编码任务上具有更好的泛化性能。最后模型的句法知识模块的引入也带来了模型效果的提升，说明利用先验知识可以改善歧义词划分不准的问题，提高新词识别准确率。通常认为在新词存在的情况下，句法分析的结果存在一定的不准确性。但实验结果的提升表明，此种不稳定的噪声的引入也对新词发现提供了一定的线索，因此依然能够对模型提供有益的信息。

4.5 模型鲁棒性分析



(a) PKU源域数据通用分词任务 (b) FR源域数据通用分词任务

(c) PKU源域数据新词发现任务 (d) FR源域数据新词发现任务

图 3: 源域数据类型对比以及目标域数据训练量趋势图

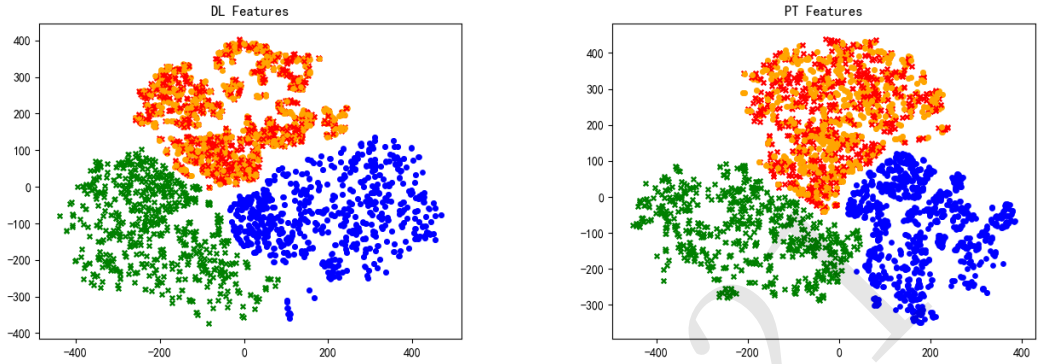
本节通过采用不同的源域类型以及目标域数据训练量进行实验分析，以验证模型鲁棒性。分别选取目标域数据集的0%，5%，20%，50%和100%的数据量构建新的目标域训练集，测试集仍使用已给出的数据。并且为了进一步验证源域数据类型是否会对新词发现的结果产生不同的影响，该实验采用不同领域数据作为源域进行测试。

如图3所示，随着目标域训练数据比例增大，在每个领域数据集上的通用分词及新词发现任务准确度都有稳定的上升趋势，在目标域训练数据比例为5%时，训练数据缺乏该领域的固有特征，模型无法进行有效特征提取，十分影响模型在领域上的分词效果。当目标域训练数据比例

增加到20%时，领域分词结果已经接近平稳，之后数据量的增多并不能带来大幅度的提升。因此只需要少量的目标域标注数据就可以有效地在专业领域进行分词及新词发现。同时也测试了使用预训练好的模型直接迁移到新的目标领域，即目标领域训练数据无标记样本参与训练的情况，观测到在所有的新领域上均取得了一个中肯的结果，因此即使在专业领域文本毫无标记语料的情况下，依然可以使用预训练好的模型进行分词及新词发现。

通过图3(a)3(c)和图3(b)3(d)对比，发现使用与DL和ZX领域相近的FR源域数据学习时，模型效果均有2%左右的提升。这也表示以后的应用中，可以选择与目标领域相同或相近的源域数据进行训练，来提升新词发现效果。

4.6 领域特征分布可视化分析



(a) DL数据集特征分布　　(b) PT数据集特征分布

图 4: 领域特征t-SNE分布图：源域特征向量、目标域特性向量、共享层双域特征向量

模型的对抗训练目的在于提取领域无关的信息，以辅助新词发现。如图4所示，本节实验使用t-SNE (Van der Maaten and Hinton, 2008)算法将源域自编码器，目标域自编码器以及共享编码器的源域和目标域共四个部分的文本特征向量投影到二维平面上进行可视化表示，进一步观测模型对领域特征学习的结果。在图4中，由叉状标记表示源域数据，圆圈标记表示目标域数据。绿色叉状表示源域编码器提取的源域数据特征向量，蓝色圆圈表示目标域编码器提取的数据特征向量；共享编码器对源域和目标域数据提取的特征向量，分别用红色叉状和橙色圆圈表示。DL领域特征图4(a)和PT领域特征图4(b)显示，独立于源域和目标域的编码器提取的特征几乎没有重叠，这说明两个域之间的数据特征存在差异，而通过共享编码器提取的源域和目标域的数据特征大致分布在一个区域，并且重叠现象明显，证明共享编码器可以有效的提取与领域无关的特征，并且这部分共享信息有利于目标专业领域通用分词及新词发现。

5 总结

本文提出了融合自编码器和对抗训练的中文新词发现模型，通过预训练自编码器融入文本语义信息；添加先验句法知识提升歧义词划分准确性；通过共享编码器融合了语义和语法信息以及跨域数据特征；最后引入对抗训练的机制，提取领域无关特征，进行新词发现。实验中使用了六个不同的专业领域数据集，以评估新词发现任务，结果显示本文方法优于现有其他方法；进行了模型析构实验，验证了模型各个模块的有效性。通过使用不同类型的源域数据和不同数量的目标域数据进行实验，验证了模型鲁棒性，最后可视化展示对抗训练方法提取到的跨域之间的相关性和差异性。在今后工作中，计划通过添加不同领域的预训练任务，融入更多的语言学知识，在极少甚至零标注样本数据中进行有效的新词发现。

参考文献

陈飞，刘奕群，魏超. 2013. 基于条件随机场方法的开放领域新词发现. 软件学报, (05):1051–1060.

Deng Cai, Hai Zhao, and Zhisong Zhang. 2017. Fast and accurate neural word segmentation for chinese. *arXiv preprint arXiv:1704.07047*.

Pengfei Cao, Yubo Chen, and Kang Liu. 2018. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 182–192.

Xinchi Chen and Xipeng Qiu. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206.

Xinchi Chen and Zhan Shi. 2017. Adversarial multi-criteria learning for chinese word segmentation. *arXiv preprint arXiv:1704.07556*.

Ning Ding, Dingkun Long, and Guangwei Xu. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. *arXiv preprint arXiv:2007.08186*.

Jingjing Gong and Xinchi Chen. 2019. Switch-lstms for multi-criteria chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6457–6464.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Shen Li and Zhe Zhao et al. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018*, pages 138–143.

Heyang Liu, PengDong Gao, and Yi Xiao. 2018a. New words discovery method based on word segmentation result. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 645–648. IEEE.

Junxin Liu, Fangzhao Wu, and Chuhan Wu. 2018b. Neural chinese word segmentation with dictionary knowledge. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 80–91. Springer.

Xipeng Qiu and Hengzhi Pei. 2019. Multi-criteria chinese word segmentation with transformer. *arXiv preprint arXiv:1906.12035*.

Likun Qiu and Yue Zhang. 2015. Word segmentation for chinese novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Georg E Schulz. 1992. Binding of nucleotides by proteins: Current opinion in structural biology. *Current opinion in structural biology*, 2(1):61–67.

Yuanhe Tian and Yan Song. 2020. Joint chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jingkang Wang, Jianing Zhou, and Jie Zhou. 2018. Multiple character embeddings for chinese word segmentation. *arXiv preprint arXiv:1808.04963*.

Jie Yang, Yue Zhang, and Shuailong Liang. 2018. Subword encoding in lattice lstm for chinese word segmentation. *arXiv preprint arXiv:1810.12594*.

Yuxiao Ye and Yue Zhang. 2019. Improving cross-domain chinese word segmentation with word embeddings. *arXiv preprint arXiv:1903.01698*.

Shuai Zhang and Lijie Wang. 2020. A practical chinese dependency parser based on a large-scale dataset. *arXiv preprint arXiv:2009.00901*.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *IJCAI*, pages 4602–4608.