基于新词发现的跨领域中文分词方法

张 军* 赖志鹏 李 学 宁更新 杨 萃
(华南理工大学电子与信息学院 广州 510641)

摘 要：深度神经网络(DNN)是目前中文分词的主流方法，但将针对某一领域训练的网络模型用于其他领域时，会因存在跨领域的未登录词(OOV)和表达鸿沟而造成性能显著下降，而在实际中对所有未知领域的训练语料进行人工标注和训练模型并不可行。为了解决这个问题，该文构建了一个基于新词发现的跨领域中文分词系统，可以自动完成从目标领域语料中提取新词、标注语料和训练网络模型的工作。此外，针对现有新词发现算法提取出的词表垃圾词串多以及自动标注语料中存在噪声样本的问题，提出了一种基于向量增强互信息和加权邻接熵的无监督新词发现算法以及一种基于对抗式训练的中文分词模型。实验中将使用北大开源新闻语料训练的网络模型提取出的特征迁移到医疗、发明专利和小说领域，结果表明该文所提方法在未登录词率、准确率、召回率和分词F值方面均优于现有方法。

关键词：中文分词；新词发现；跨领域；向量增强互信息；对抗式训练
中图分类号：TP931 **文献标识码**：A **文章编号**：1009-5896(2022)09-3241-08
DOI: 10.11999/JEIT210675

Cross-domain Chinese Word Segmentation Based on New Word Discovery

ZHANG Jun LAI Zhipeng LI Xue NING Gengxin YANG Cui
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou* 510641, *China*)

Abstract: Deep Neural Network (DNN) is the major method in current Chinese word segmentation. However, its performance is significantly degraded when the network trained for one domain is used in other domains due to the Out Of Vocabulary (OOV) words and expression gaps. In this paper, a cross domain Chinese word segmentation system based on new word discovery is built to handle the OOV word and expression gap problems. An unsupervised new word discovery algorithm based on vector enhanced mutual information and weighted adjacency entropy, and a Chinese word segmentation model based on adversarial training are also proposed to improve the performance of the baseline system. Experimental results show that the proposed method is superior to the conventional methods in the OOV rates, precisions, recalls and F-scores.

Key words: Chinese word segmentation; New word discovery; Cross-domain; Vector enhancement mutual information; Adversarial training

1 引言

词语是中文文本中包含语义信息并且能够独立使用的最小结构单元，因此中文分词(Chinese Word Segmentation, CWS)是中文自然语言处理(Natural Language Processing, NLP)的基础，其性能好坏将对NLP下游任务的效果产生直接影响。

早期的中文分词方法主要包括机械分词法[1]和统计分词法[2,3]两种。机械分词法需要预先构造一个足够大的中文词表，然后通过设置词表中词语的组合规则来对中文句子进行切分。统计分词法则是根据相邻字之间的共现频率来计算它们构成词语的可信度，无需预先构建词表。由于这两种方法所使用的分词模型都较为简单，不能很好地描述复杂的中文构词规律，因此分词性能并不理想。随着深度学习技术的提出和发展，近年来提出了多种利用深度神经网络来进行中文分词的方法[4,5]。与传统的

收稿日期：2021-07-06；改回日期：2021-09-14；网络出版：2021-12-25
*通信作者： 张军 eejzhang@scut.edu.cn
基金项目：国家自然科学基金(61871191)，广东省自然科学基金(2020A1515010962)，广州市科技计划(202002030251)
Foundation Items: The National Natural Science Foundation of China (61871191), The Natural Science Foundation of Guangdong Province (2020A1515010962), The Natural Science Foundation of Guangzhou (202002030251)

分词方法不同，基于深度神经网络的分词方法将中文分词当成序列标注任务，以人工标注的数据集来训练网络，在无需获取中文词表和人为构造规则，也不需要人为构造特征模板的情况下，利用深度网络的强大建模能力，能获得远高于传统方法的准确率(Precision)和召回率(Recall)，因此成为目前中文分词的主流技术。

在基于深度神经网络的中文分词方法中，首先需要使用大量标注好的语料来训练网络模型，然后利用训练好的网络模型对无标注的测试语料进行分词。当训练语料的领域(源领域)与测试语料的领域(目标领域)属于同一个领域时，这种方法能通常能取得很好的效果，但当源领域和目标领域不属于同一个领域，即跨领域(cross domain)时，其性能将会显著降低。造成这种现象的原因主要有两个，一是未登录词(Out Of Vocabulary, OOV)，即目标领域中存在着大量未在源领域中出现过的词语，这些词语对于网络模型来说是未知样本，难以正确识别。另一个原因是领域之间的表达鸿沟，即不同领域的语言表达是有差异的，使得网络模型在源领域上学习的特征对于目标领域并不具有很好的泛化性能。解决未登录词和表达鸿沟最简单的方法是对目标领域的语料进行标注并重新训练模型，但由于在实际中对所有未知领域的训练语料进行人工标注需要非常高的成本，并不可行，因此如何有效地解决中文分词系统的领域适应性，特别是未登录词和表达鸿沟问题，是目前中文分词的最大难点之一。现有的研究中，在模型的训练中结合词典或字/词向量是解决未登录词的最常用的方法[6]，而迁移学习则是解决表达鸿沟的主要方法[7]。尽管目前对跨领域中文分词中的未登录词或表达鸿沟问题已有一定的研究，但现有的文献所提方法大多只针对两者之一，而同时解决两个问题的研究成果尚不多见。

本文针对跨领域中文分词中的未登录词和表达鸿沟问题，首先采用现有技术构建了一个基于新词发现的跨领域中文分词系统，实现了自动从目标领域语料中提取新词、标注语料和训练网络模型的功能。然后针对现有新词发现算法提取出的词表垃圾词串多的缺点，提出了一种基于向量增强互信息和加权邻接熵的无监督新词发现算法，以提高新词词表提取的准确率和领域性。最后，针对自动标注语料中存在噪声样本的不足，提出了一种基于对抗式训练的中文分词模型，有效提高了分词网络模型训练的鲁棒性。

文章的其余部分组织如下：第2节介绍了本文搭建的基线系统，第3节提出了基于向量增强互信息和加权邻接熵的无监督新词发现算法，第4节提出了基于对抗式训练的中文分词算法，第5节是实验结果和分析，最后一节给出了结论。

2 基线系统

为了同时处理跨领域分词中的未登录词和表达鸿沟问题，本文构建的基线系统包含新词发现、自动标注和跨领域分词3个部分，结构如图1所示。首先使用新词发现算法从各个目标领域语料中提取出该领域的新词词表，然后利用该新词词表对无标注的目标领域语料进行自动标注，以降低目标领域语料的未登录词率，最后使用自动标注好的语料训练分词模型，并使用该模型来对目标领域进行分词。在这个系统中，新词发现能显著减少跨领域分词中的未登录词率，而对目标领域语料的自动标注并在此基础上训练适用于目标领域的分词模型，则能有效解决跨领域分词中的表达鸿沟问题。

新词发现包含语料预处理、候选词提取和候选词过滤3个步骤。目标领域的中文语料首先按照非中文字符的方式进行切割，并剔除非汉字字符，然后使用N-Gram的方法[8]从目标领域语料中提取出所有的候选字符串，得到候选词集。此时得到的候选词集既包含了正确的词语，又包含了大量错误的字符组合，因此需要对词集中的词进行筛选。互信息(Mutual Information, MI)和邻接熵(Branch Entropy, BE)相结合的方法是目前最常用的词集筛选方法[9]，首先统计每个候选词在目标领域语料中的词频后，然后采用下式计算出每个词的得分

$$\text{score}(\boldsymbol{w}) = \text{MI}(\boldsymbol{w}) + \min\left(H_{\text{l}}(\boldsymbol{w}), H_{\text{r}}(w)\right) \tag{1}$$

其中

$$\text{MI}(\boldsymbol{w}) = \min_{1\le i<n} \ln \frac{p(\boldsymbol{w})}{p(c_1c_2\cdots c_i)\cdot p(c_{i+1}\cdots c_n)} \tag{2}$$

为互信息，$\boldsymbol{w}=c_1c_2\cdots c_ic_{i+1}\cdots c_n$为长度为$n$的候选词，$c_i$为$\boldsymbol{w}$中的第$i$个字符，$p(\cdot)$表示在语料中出现的概率。$H_{\text{l}}(\boldsymbol{w})$和$H_{\text{r}}(\boldsymbol{w})$分别为候选词$\boldsymbol{w}$的左邻接熵和右邻接熵，定义为

$$H_{\text{l}}(\boldsymbol{w}) = -\sum_{i=1}^{k} p(l_i)\cdot\ln\left(p(l_i)\right) \tag{3}$$

$$H_{\text{r}}(\boldsymbol{w}) = -\sum_{i=1}^{k'} p(r_i)\cdot\ln\left(p(r_i)\right) \tag{4}$$

其中

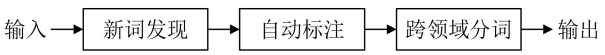

图 1 基线系统的结构

$$p(l_i)=\frac{N(l_i)}{\sum_{j=1}^{k}N(l_j)} \tag{5}$$

$$p(r_i)=\frac{N(r_i)}{\sum_{j=1}^{k'}N(r_j)} \tag{6}$$

$p(l_i)$表示左邻接字l_i出现的概率，$N(l_i)$和$N(l_j)$为字符l_i和l_j在字片段w左侧出现的次数，$p(r_i)$表示右邻接字r_i出现的概率，$N(r_i)$和$N(l_j)$为字符r_i和r_j在字片段w右侧出现的次数，k和k'分别表示左邻接字和右邻接字的数量。过滤掉得分较低的候选词，并剔除源领域词表中的词和中文常用词，即可得到一个干净的目标领域新词词表。

自动标注中，首先根据目标领域的新词词表使用逆向最大匹配算法(Backward Maximum Matching, BMM)[10]对目标领域语料进行初步的切分，然后利用有标注的源领域语料训练分词模型，并使用该模型对目标领域的语料进行完全切分，得到自动标注的目标领域语料。分词模型是中文分词系统的核心，由于目前基于深度神经网络的分词方法均将中文分词当成序列标注任务，因此主流的中文分词方法是使用双向长短时记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)[11]加上条件随机场模型(Conditional Random Fields, CRF)[12]。由于在BiLSTM中，输入之间相互依赖使得模型在处理当前字符时还可以提取到上下文里的语境和语义信息，并且从理论上来说这个上下文可以扩展到全文，而CRF模型属于统计模型，可以在分词模型中加入有关于语料的统计信息，能很好地弥补深度模型无法提取浅层特征的劣势，因此BiLSTM+CRF在领域内分词和跨领域分词中都取得了很好的效果。但BiLSTM在实际使用中存在着训练速度慢、信息冗余，在获取远距离依赖时容易出现梯度爆炸或者梯度弥散的缺点。为此，本文的基线系统中使用了门控卷积神经网络(Gated Convolutional Neural Network, GCNN)[13]来代替BiLSTM。GCNN是带有线性门控的卷积神经网络，使用线性门控后能令模型在堆叠获取远距离上下文时可以遗忘不重要的信息而只保留重要的信息，远距离依赖效果将会变得更好，不仅可以进一步降低梯度弥散的现象，还可以保留CNN的非线性能力。

跨领域分词时，利用自动标注好的目标领域语料训练出一个适用于目标领域的GCNN-CRF模型，即可以使用该模型对目标领域内的测试语料进行分词。由于该模型是使用自动标注好的目标领域语料训练出来的，因此能克服跨领域中文分词的未登录词和表达鸿沟问题。

3　基于向量增强互信息和加权邻接熵的无监督新词发现算法

传统基于MI+BE的无监督新词发现算法认为互信息可以表示字符串内部聚合度的大小，左右邻接熵可以表示字符串边界自由度的高低，因此将互信息和邻接熵直接相加可以同时衡量字符串内部聚合度和边界自由度的高低。但在实验中发现，使用MI+BE算法提取的新词词表中存在大量垃圾词串，例如“过程中”、“线城市”等非词语的固定搭配由于具有较大的词频和互信息，并且邻接熵也较大，很容易被错误地认为是一个合理的新词。究其原因，MI+BE算法一方面在判定内部凝结度上只利用了语料中的统计信息，使得一些常用搭配因为凝固度较高而被认为也是新词，另一方面在判定边界自由度上只利用了左右邻接熵中的较小值，使得一些错误词串也被认为是新词，造成提取出的新词词表中含有较多的垃圾词串。本文针对MI+BE算法的不足，提出了基于向量增强互信息和加权邻接熵的无监督新词发现算法。

3.1　向量增强互信息

对于一个新词而言，它内部的片段应该是紧密结合并且很大概率是一起出现在句子中的，也就是说这些片段之间必然就会有着相似的上下文语境，因此使用上下文语境的相关性来进一步描述字符串内部的结合程度对新词发现应有一定的帮助。本文借助基于语义的词向量来对互信息进行改进，提出向量增强互信息(Vector Enhancement Mutual Information, VEMI)的概念。

假设将候选词$\boldsymbol{w}=c_1c_2\cdots c_ic_{i+1}\cdots c_n$分成两个字符片段$c_1c_2\cdots c_i$和$c_{i+1}\cdots c_n$，其向量表达为$\mathbf{Vec}_{c_1c_2\cdots c_i}=(a_1,a_2,\cdots,a_n)$和$\mathbf{Vec}_{c_{i+1}\cdots c_n}=(b_1,b_2,\cdots,b_n)$，定义它们的余弦相似度为

$$\mathrm{simCos}=\frac{\sum_{i=1}^{n}a_i\cdot b_i}{\sqrt{\left(\sum_{i=1}^{n}a_i^2\right)\left(\sum_{i=1}^{n}b_i^2\right)}} \tag{7}$$

将余弦相似度作为相关性的另一个计算因素加入到$\mathrm{MI}(\boldsymbol{w})$中，得到向量增强互信息为

$$\mathrm{VEMI}(\boldsymbol{w})=\min_{1\le i<n}\ln\frac{p(\boldsymbol{w})}{p(c_1c_2\cdots c_i)\,p(c_{i+1}\cdots c_n)}+\alpha\cdot\mathrm{simCos} \tag{8}$$

其中，α是一个超参数，用于平衡互信息和语义相似性的重要性。在VEMI的计算中加入字符片段的余弦相似度，使其输出不但受字符片段结合的紧密

度影响，还考虑了上下文的相关性，因此应当优于传统的MI准则。

由于字符片段$c_1c_2\cdots c_i$和$c_{i+1}\cdots c_n$不一定是一个合理的词，在实际中难以直接通过常规的词向量训练得到它们的向量$\mathbf{Vec}_{c_1c_2\cdots c_i}$和$\mathbf{Vec}_{c_{i+1}\cdots c_n}$，因此本文采用的方法是先将语料切分为字符，然后使用Word2Vec[14]模型进行训练得到所有字符对应的字向量，这些字向量中即包含有一定范围内的上下文语义信息，最后将得到的字向量进行相加求平均得到字符片段的向量。设任意字符c_j通过Word2Vec训练得到的字向量为$\mathbf{Vec}_{c_j}=(a_{j1},a_{j2},\cdots,a_{jn})$，则字符片段$c_1c_2\cdots c_i$的向量采用下式计算

$$\mathbf{Vec}_{c_1c_2\cdots c_i}=\frac{1}{i}\sum_{j=1}^{i}\mathbf{Vec}_{c_j} \tag{9}$$

3.2 加权邻接熵

根据式(3)和式(4)计算得到字符片段的左右邻接熵后，传统的方法是选择较小的熵作为指标来对字片段的边界进行衡量，这种方式虽然简单，但是并没有充分考虑到左右两边的邻接熵信息，在很多情况下是不合适的，例如候选词语“红皮病”，由于其在文中多是单独成句子出现，因此其左邻接熵很低，传统的算法会将这个词语剔除，但“红皮病”在文本中却是一个新词。为此，本文对传统的邻接熵进行了改进，采用加权的方式来同时利用左右两边的邻接熵信息，加权邻接熵的计算为

$$\begin{aligned}\mathrm{BE}(\boldsymbol{w})=&H_{\mathrm{l}}(\boldsymbol{w})\cdot\ln\frac{H_{\mathrm{r}}(\boldsymbol{w})}{|H_{\mathrm{l}}(\boldsymbol{w})-H_{\mathrm{r}}(\boldsymbol{w})|+\varepsilon}\\&+H_{\mathrm{r}}(\boldsymbol{w})\cdot\ln\frac{H_{\mathrm{l}}(\boldsymbol{w})}{|H_{\mathrm{l}}(\boldsymbol{w})-H_{\mathrm{r}}(\boldsymbol{w})|+\varepsilon}\end{aligned} \tag{10}$$

其中，$\mathrm{BE}(\boldsymbol{w})$表示加权后的邻接熵，$H_{\mathrm{l}}(\boldsymbol{w})$和$H_{\mathrm{r}}(\boldsymbol{w})$分别为词$\boldsymbol{w}$的左邻接熵和右邻接熵，$\varepsilon$为一个小的正数。式(10)的对数部分相当于对左右邻接熵加上一个权重，其作用主要有两个方面，一是令较大的熵权值变小，较小的熵权值变大，从而使最终结果不再仅由其中的较小值所支配；二是当一个字符串片段的左右邻接熵都比较小，但相差不大时，有很大可能是一个合理的词，式(10)会增加这种情况下左右熵的权重，使得总得分变大。以“关节病”这个片段为例，实验中统计得到其左邻字分别有{疗，对}，出现次数依次为{1，1}，右邻字分别有{人，或，的}，出现的次数依次为{8, 1, 1}。据统计得到的左右邻字及其出现次数，可以得到其左右邻接熵分别为0.693和0.639，选择其中较小的右邻熵作为得分，则得分过小，“关节病”这个片段将不会认为是一个合理的新词。而根据式(10)可以计算得到其邻接熵为3.341，这是一个较大的熵值，会认为这个片段是一个合理的新词。由此可见，使用加权邻接熵比直接使用左右邻接熵中的较小值进行判断效果更好。

3.3 候选词筛选

通过N-Gram方法切分语料得到了所有长度不大于预设值的候选字符串片段，并使用Word2Vec训练目标领域语料中的汉字，得到了目标领域语料中所有汉字的字向量后，采用向量增强互信息和加权邻接熵来对所有的候选词$\boldsymbol{w}$进行打分

$$\mathrm{score}(\boldsymbol{w})=\sigma\left(a\cdot N\left(\mathrm{VEMI}(\boldsymbol{w})\right)+b\cdot N\left(\mathrm{BE}(\boldsymbol{w})\right)\right) \tag{11}$$

其中，σ为sigmoid函数，$N(\cdot)$表示最大最小值归一化

$$N(f(\boldsymbol{w}))=\frac{f_{\max}-f(w)}{f_{\max}-f_{\min}} \tag{12}$$

$f_{\max}$和$f_{\min}$为$f(\boldsymbol{w})$的最大和最小值，系数a和b分别表示互信息和邻接熵在最终得分中所占据的权重大小。计算出候选词的得分后，剔除得分小于阈值的候选词。

经过得分筛选后，第2步将进行词频筛选。一个片段如果是一个合理的词语，那么这个片段在语料必然是多次出现的，本文将词频的最小值设定为8，出现次数小于8的片段即使得分较大也不认为是一个合理的词，将其从候选词中删除。

通过观察发现，采用以上步骤得到的新词词表中仍然存在少量诸如“导致了”、“扩展到”等错误词语，因此本文对词表进行了第3次筛选：统计候选词的首字和尾字的出现次数，如果这些字出现的次数大于一定值就认为这些字构成的词属于常用搭配而不是新词，比如“了”就在首尾中出现了261次，高于预设的阈值100，那么认为这些词语就是不合理的新词将其进行删除。

需要注意的是，由于分词只是将合理的词切分出来，不涉及词语语义的理解，因此中文的同义词只要是正确的词，组成它们的字之间的互信息以及它们与邻近字之间的联系与其他正确词语具有相似的特性，同样可以采用本文方法进行切分，无需特殊处理。

4 基于对抗式训练的中文分词模型

尽管本文提出的新词发现算法提取出的新词词表具有较高的准确性和领域性，但目标领域语料是完全基于新词词表和分词算法进行自动标注的。由于词表和分词算法本身并不能保证完全正确，因此自动标注的语料会存在着一定数量的噪声样本。基线系统中使用的GCNN-CRF算法原本是基于正确

标注好语料而设计的，并未考虑到训练语料中会存在噪声，因此并不具有抑制噪声对模型影响的能力，自动标注语料中的噪声将会影响分词模型的性能。针对这个问题，本文提出了一种基于对抗式训练的中文分词模型，通过单独提取出源领域和目标领域的共有特征来提高目标领域特征的鲁棒性，其结构如图2所示。

由图2可以看到，本文提出的跨领域分词模型包含3个GCNN编码器，分别是源领域GCNN编码器、目标领域GCNN编码器和共享GCNN编码器。源领域编码器和目标领域编码器只接收各自领域的文本作为输入，用于提取各自领域独有特征，共享编码器则同时接收两个领域的文本作为输入，提取两个领域的共有特征。源领域编码器得到的独有特征和共享编码器得到的共有特征组合即可得到源领域的文本特征，再将这个特征输入到CRF中对源领域的文本预测词位标签。目标领域的处理方式与源领域相同。共享编码器的目标是尽可能提取出源领域和目标领域共有的特征，文中采用了对抗式训练来对其进行优化，将共享编码器中提取出的共有特征输入到一个文本判别器TextCNN[15]中，使用文本判别器来判别共享编码器输出的特征是来自源领域还是目标领域。

模型训练时，上支路和中间支路、下支路和中间支路的训练交替进行，即源领域编码器和目标领域编码器的参数交替更新，共享编码器和文本判别器的参数每次迭代都更新。假设句子$\boldsymbol{s}_{\mathrm{s}}$和$\boldsymbol{s}_{\mathrm{t}}$分别来自源领域和目标领域，它们通过源领域编码器、目标领域编码器和共享编码器后分别得到源领域的独有特征$\boldsymbol{H}_{\mathrm{s}}$、目标领域的独有特征$\boldsymbol{H}_{\mathrm{t}}$和两个领域的共有特征$\boldsymbol{H}^{*}$。$\boldsymbol{H}^{*}$提取越准确，其领域性就越低，文本判别器就越难通过它判别句子来自哪个领域，可以据此对共享编码器和文本判别器进行对抗式训练。将$\boldsymbol{H}^{*}$输入文本判别器的TextCNN，TextCNN将产生一个$\boldsymbol{H}^{*}$属于源领域或目标领域的概率，令$G(H^{*})$表示TextCNN的输出，文本判别器总损失函数设置为

$$L_{\mathrm{s}}^{*}=-\mathrm{E}_{\mathrm{s}}\left[\lg\left(G\left(\boldsymbol{H}^{*}\right)\right)\right]-\mathrm{E}_{\mathrm{s}}\left[\lg\left(1-G\left(\boldsymbol{H}^{*}\right)\right)\right] \tag{13}$$

$$L_{\mathrm{t}}^{*}=-\mathrm{E}_{\mathrm{t}}\left[\lg\left(G\left(\boldsymbol{H}^{*}\right)\right)\right]-\mathrm{E}_{\mathrm{t}}\left[\lg\left(1-G\left(\boldsymbol{H}^{*}\right)\right)\right] \tag{14}$$

其中，L_{s}^{*}为文本判别器在上支路和中间支路训练时的损失函数，$\mathrm{E}_{s}[\cdot]$表示对源领域的句子求期望，L_{t}^{*}为文本判别器在下支路和中间支路训练时的损失函数，$\mathrm{E}_{\mathrm{t}}[\cdot]$表示对目标领域的句子求期望。源领域和目标领域的总损失函数$L_{\mathrm{s}}^{\mathrm{total}}$和$L_{\mathrm{t}}^{\mathrm{total}}$分别设置为

$$L_{\mathrm{s}}^{\mathrm{total}}=L_{\mathrm{s}}+\alpha\cdot L_{\mathrm{t}}+L_{\mathrm{s}}^{*} \tag{15}$$

$$L_{\mathrm{t}}^{\mathrm{total}}=L_{\mathrm{s}}+\alpha\cdot L_{\mathrm{t}}+L_{\mathrm{t}}^{*} \tag{16}$$

其中，L_{s}和L_{t}分别为源领域编码器和目标领域编码器GCNN的交叉熵损失函数，α为权重因子。考虑到目标领域中的文本是自动标注的，其中有不正确标注的样本，目标领域中的样本在训练中的置信度应该低于源领域，因此设置$0<\alpha<1$。

通过共享编码器和文本判别器的对抗式训练，可以使共享编码器提取的特征包含更少的源领域和目标领域的独有特征，越来越接近两个领域的共有特征。与单个GCNN-CRF相比，本文方法有以下优势：(1)由于源领域的语料是正确标注的，因此两个领域的共有特征在理想情况下不含噪声，目标领域中的标注噪声只存在于其独有特征中。将目标领域的共有特征和独有特征分离，可以将噪声的影响限制在一定范围内，从而提高目标领域分词对标注错误的鲁棒性。(2)将源领域损失与目标领域损失的和作为总损失，与原GCNN-CRF模型相比相当于在训练过程中加入了正则化，可以起到防止过拟合和增强鲁棒性的作用。

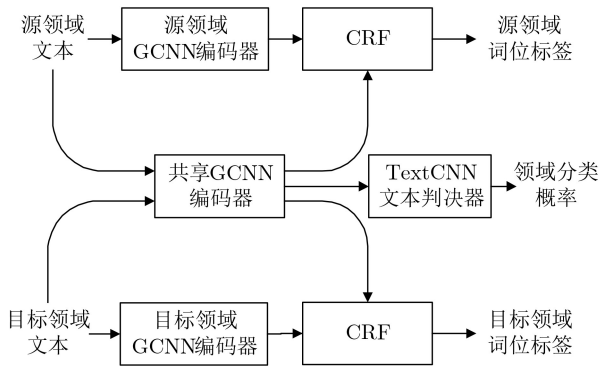

图 2 基于对抗式训练的中文分词模型

5　实验结果及分析

5.1　实验设置

实验中采用的数据分为源领域和目标领域两个部分，其中源领域数据为中文分词领域中普遍使用的北大开源新闻语料[16]，目标领域数据包括医疗、小说《诛仙》和《斗罗》、发明专利3个领域的语料，这些目标领域语料中都随机选取一部分做了人工标注作为测试集，其中训练集和测试集的比例大致为5:1。各个数据集的大小如表1所示。

实验中所使用的深度神经网络的训练和识别均基于开源框架tensorflow1.14，所有数据的编码格式为UTF-8，GCNN网络维度为200，层数为5，

Dropout率为0.2，学习率为0.001，Epoch数为15。新词发现中的使用N-Gram方法切分的字符串片段最大长度为6，候选词得分阈值为0.95，词频阈值为8，首字和尾字的出现次数阈值为100。以上阈值均为经验阈值，通过观察和实验来确定。字符串最大长度、词频阈值、首字和尾字出现次数阈值分别通过观察统计训练语料分词结果中正确词语的长度、词频、典型的首字和尾字(如上文提到的“了”字)出现次数得到，这些阈值设置过严容易导致正确的词语被切分开或丢弃，而设置过宽松则容易产生错误的字词组合。候选词得分阈值则是在实验中尝试多个阈值后，选取未登录词率最优的0.95。

实验中对新词发现和中文分词算法的性能采用了不同的评价指标。新词发现算法的主要目的是用于解决跨领域中文分词中的未登录词问题，因此实验中使用未登录词率(即未登录词数量与总词数的比值)来作为评价指标。中文分词算法的性能则采用了准确率、召回率和F值(F-measure)3个常用的评价指标来衡量。

5.2 新词发现

为了衡量本文提出的新词发现算法的性能，实验中首先分别使用MI+BE算法和本文提出的新词发现算法从目标领域训练语料上提取该领域相对源领域语料独有的新词，再利用新词词表对目标领域测试语料进行自动标注，并统计标注过程中出现的未登录词占总词数的比例。实验中还对不进行新词发现、直接使用源领域词表对目标领域测试语料进行自动标注时的未登录词率进行了统计。表2给出了无新词发现、MI+BE算法和本文提出的新词发现算法应用于目标领域测试语料时的未登录词率。

表 1 实验中使用的语料大小(Byte)

语料	句子(k)	词语(M)	字符(M)
新闻	53.7	1.3	2.1
医疗	32.0	0.7	1.2
《诛仙》	59.0	2.1	3.0
《斗罗》	40.0	2.0	0.9
发明专利	17.0	0.6	0.9

由表2的结果可以看出，使用了新词发现算法的未登录词率比无新词发现、直接使用源领域词表时有显著的下降，同时，本文所提的新词发现算法要明显优于传统的MI+BE算法，在各个语料上都取得了最好的效果，说明了本文方法的有效性。

为了更好地检验本文算法所提取的新词的合理性，表3给出了MI+BE算法和本文算法从各个语料中提取的前20个最频繁出现词中垃圾词串的数目。由表3可以看到，本文方法提取的词表更准确，有效地减少了词表中无意义的垃圾词串数量。

5.3 基于对抗式训练的分词算法

为了测试本文提出的基于对抗式训练的分词算法的效果，表4给出了GCNN-CRF与本文对抗式训练模型在目标领域测试语料上分词的准确率、召回率和F值，其中基线系统使用了MI+BE的新词发现算法和GCNN-CRF分词算法，GCNN-CRF使用了本文的新词发现算法和GCNN-CRF分词算法，本文方法使用了本文的新词发现算法和对抗式训练模型。由表4可以看到，基线系统的性能最差，使用了本文新词发现算法的GCNN-CRF性能次之，本文方法性能最优，这说明：(1)由于传统的MI+BE算法提取的新词词表中存在着较多的缺失和错误，本文的新词发现算法能更准确地提取新词，因此使用了本文新词发现算法的GCNN-CRF性能显著优于基线系统。(2)由表3可知，本文的新词发现算法中仍存在着少量标注错误的噪声样本，而在中

表 2 不同方法的未登录词率(%)

算法	无新词发现	MI+BE	本文方法
医疗	25.93	16.31	**5.42**
《诛仙》	15.52	8.24	**1.43**
《斗罗》	11.15	7.06	**1.23**
发明专利	18.39	11.27	**3.45**

表 3 前20个最频繁出现词中垃圾词串数(个)

语料	医疗	《诛仙》	《斗罗》	发明专利
MI+BE	4	5	7	6
本文方法	1	1	2	6

表 4 基于对抗式训练的分词算法效果

性能指标	准确率(%)			召回率(%)			F值		
算法	基线	GCNN_CRF	本文方法	基线	GCNN_CRF	本文方法	基线	GCNN_CRF	本文方法
医疗	71.7	80.7	**82.4**	74.3	82.0	**83.7**	73.0	81.4	**83.0**
《诛仙》	77.8	89.3	**90.3**	75.6	87.5	**87.7**	76.7	88.4	**89.0**
《斗罗》	81.7	92.1	**92.8**	81.7	91.9	**92.4**	81.0	92.0	**92.6**
发明专利	84.3	88.1	**89.8**	81.6	87.1	**87.2**	82.9	87.6	**88.5**

文分词中引入对抗式训练可以有效地降低噪声样本对模型的影响，使模型在跨领域分词时取得比传统GCNN-CRF更高的准确率。

5.4 与现有方法的对比

为了衡量本文方法的整体性能，实验中将本文方法与文献[6]所提的方法进行了对比。文献[6]提出的分词模型首先采用人工识别的方法提前获得目标领域词典，然后将该领域词典作为先验知识和源领域标注语料组成训练集，通过训练改进的BiLSTM+CRF网络模型实现跨领域分词。由于本文所用的目标领域语料没有现成的人工标注词典，因此文献[6]的方法中人工词典使用本文的新词发现算法构造的词典代替。从表5可以看到，本文方法的性能显著优于文献[6]的方法。

表 5　本文方法与现有方法的性能对比

性能指标	准确率(%)			召回率(%)			F值		
算法	基线系统	文献[6]	本文方法	基线系统	文献[6]	本文方法	基线系统	文献[6]	本文方法
医疗	71.7	80.1	**82.4**	74.3	82.3	**83.7**	73.0	81.2	**83.0**
《诛仙》	77.8	86.7	**90.3**	75.6	**88.9**	87.7	76.7	87.8	**89.0**
《斗罗》	81.7	91.9	**92.8**	81.7	92.1	**92.4**	81.0	92.0	**92.6**
发明专利	84.3	85.5	**89.8**	81.6	86.3	**87.2**	82.9	85.9	**88.5**

6　结束语

未登录词和表达鸿沟是目前跨领域中文分词中的难点问题，而目前同时解决两个问题的研究尚不多见。本文针对这两个问题，构建了一个基于新词发现的跨领域中文分词系统，可以自动完成从目标领域语料中提取新词、标注语料和训练网络模型的工作。在此基础上，针对常用的MI+BE新词发现算法提取出的词表垃圾词串多的问题，对互信息和邻接熵的提取进行了改进，提出了一种基于向量增强互信息和加权邻接熵的无监督新词发现算法；针对自动标注语料中存在的噪声文本问题，提出了一种基于对抗式训练的中文分词模型，使用对抗式训练来提取源领域和目标领域的共有特征，以提高中文分词系统的鲁棒性和跨领域表达能力。实验中将使用北大开源新闻语料训练的网络模型提取出的特征迁移到医疗、发明专利和小说领域，结果表明本文所提方法在未登录词率、准确率、召回率和分词F值方面均优于现有模型。

参 考 文 献

[1] 陈平, 刘晓霞, 李亚军. 基于字典和统计的分词方法[J]. 计算机工程与应用, 2008, 44(10): 144–146. doi: 10.3778/j.issn.1002-8331.2008.10.042.
CHEN Ping, LIU Xiaoxia, and LI Yajun. Chinese word segmentation based on dictionary and statistics[J]. *Computer Engineering and Applications*, 2008, 44(10): 144–146. doi: 10.3778/j.issn.1002-8331.2008.10.042.

[2] WU Andi and JIANG Zixin. Word segmentation in sentence analysis[C]. 1998 International Conference on Chinese Information Processing, Beijing, China, 1998: 169–180.

[3] 朱聪慧, 赵铁军, 郑德权. 基于无向图序列标注模型的中文分词词性标注一体化系统[J]. 电子与信息学报, 2010, 32(3): 700–704. doi: 10.3724/SP.J.1146.2009.00214.
ZHU Conghui, ZHAO Tiejun, and ZHENG Dequan. Joint Chinese word segmentation and POS tagging system with undirected graphical models[J]. *Journal of Electronics & Information Technology*, 2010, 32(3): 700–704. doi: 10.3724/SP.J.1146.2009.00214.

[4] YUAN Zheng, LIU Yuanhao, YIN Qiuyang, *et al*. Unsupervised multi-granular Chinese word segmentation and term discovery via graph partition[J]. *Journal of Biomedical Informatics*, 2020, 110: 103542. doi: 10.1016/j.jbi.2020.103542.

[5] DU Jinlian, MI Wei, and DU Xiaolin. Chinese word segmentation in electronic medical record text via graph neural network-bidirectional LSTM-CRF model[C]. 2020 IEEE International Conference on Bioinformatics and Biomedicine, Seoul, Korea, 2020: 985–989. doi: 10.1109/BIBM49941.2020.9313165.

[6] WANG Qi, ZHOU Yangming, RUAN Tong, *et al*. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. *Journal of Biomedical Informatics*, 2019, 92: 103133. doi: 10.1016/j.jbi.2019.103133.

[7] XU Jingjing, MA Shuming, ZHANG Yi, *et al*. Transfer deep learning for low-resource Chinese word segmentation with a novel neural network[C]. The 6th National CCF Conference on Natural Language Processing and Chinese Computing, Dalian, China, 2017: 721–730. doi: 10.1007/978-3-319-73618-1_62.

[8] BELLEGARDA J R. Statistical language model adaptation:

Review and perspectives[J]. *Speech Communication*, 2004, 42(1): 93–108. doi: 10.1016/j.specom.2003.08.002.

[9] 刘伟童, 刘培玉, 刘文锋, 等. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究, 2019, 36(5): 1293–1296. doi: 10.19734/j.issn.1001-3695.2017.11.0745.
LIU Weitong, LIU Peiyu, LIU Wenfeng, *et al*. New word discovery algorithm based on mutual information and branch entropy[J]. *Application Research of Computers*, 2019, 36(5): 1293–1296. doi: 10.19734/j.issn.1001-3695.2017.11.0745.

[10] 罗桂琼, 费洪晓, 戴弋. 基于反序词典的中文分词技术研究[J]. 计算机技术与发展, 2008, 18(1): 80–83. doi: 10.3969/j.issn.1673-629X.2008.01.022.
LUO Guiqiong, FEI Hongxiao, and DAI Yi. Research of Chinese segmentation based on converse segmentation dictionary[J]. *Computer Technology and Development*, 2008, 18(1): 80–83. doi: 10.3969/j.issn.1673-629X.2008.01.022.

[11] YAO Yushi and HUANG Zheng. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]. The 23rd International Conference on Neural Information Processing, Kyoto, Japan, 2016: 345–353. doi: 10.1007/978-3-319-46681-1_42.

[12] LIU Liyuan, SHANG Jingbo, REN Xiang, *et al*. Empower sequence labeling with task-aware neural language model[C]. The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, United States, 2018.

[13] KAN Zhigang, QIAO Linbo, YANG Sen, *et al*. Event arguments extraction via dilate gated convolutional neural network with enhanced local features[J]. *IEEE Access*, 2020, 8: 123483–123491. doi: 10.1109/ACCESS.2020.3004378.

[14] MIKOLOV T, CHEN Kai, CORRADO G, *et al*. Efficient estimation of word representations in vector space[C]. The 1st International Conference on Learning Representations, Scottsdale, Arizona, 2013.

[15] KIM Y. Convolutional neural networks for sentence classification[C]. The 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 1746–1751. doi: 10.3115/v1/D14-1181.

[16] Beijing Universty, City University of Hong Kong, CKIP, *et al*. The second international Chinese word segmentation bakeoff data[EB/OL]. http://sighan.cs.uchicago.edu/bakeoff2005/, 2005.

张 军：男，副教授，研究方向为语音信号处理、水声信号处理.
赖志鹏：男，硕士生，研究方向为语音信号处理、自然语言处理.
李 学：男，硕士，研究方向为自然语言处理.
宁更新：男，副教授，研究方向为语音信号处理、水声信号处理.
杨 萃：女，副教授，研究方向为信号处理、超声机器人.

责任编辑：马秀强