文章编号：1003-0077（2011）00-0000-00

基于神经网络纠正器的领域分词方法*

吴佳林[1]，唐晋韬[1]，李莎莎[1]，王挺[1]

（1.国防科技大学，湖南 长沙 410073）

摘要：本文提出了一种基于神经网络的中文分词方法，以提高分词系统向新领域迁移的适应性和灵活性。本文方法采用了对现有分词器分词结果进行纠正的思路。这种基于纠正的两阶段方法与分词模型解耦，避免了对源领域语料和分词器构建方式的依赖。然而现有的基于纠正的方法依赖于特征工程，无法自动适应不同领域。本文利用神经网络对纠正器进行建模，在无需手工设计特征的情况下即可实现领域适应。实验表明，与当前方法相比，文本方法在领域文本上具有更好的分词性能和鲁棒性，尤其在未登录词召回率方面提升显著。

关键词：中文分词；领域适应；神经网络

中图分类号：TP391　　**文献标识码：**A

Domain Adaptation for Chinese Word Segmentation based on Neural Network Corrector

Jialin Wu[1], Jintao Tang[1], Shasha Li[1], Ting Wang[1]

(1. National University of Defense Technology, Changsha, Hunan 410073, China)

Abstract: This paper proposes a neural network based method for Chinese Word Segmentation to enhance its adaptability and flexibility when transformed to a new domain. Our method bases on the idea of revising the results of an existing segmenter. This two-phase correction based method does not depend on both the source domain data and the way of building a segmenter. However, the existing method based on the correction relies on the feature engineering, which is hard to automatically adapt different domains. We propose a neural network based corrector to conduct the domain adaptation, which does not require any hand-crafted features. Experimental results show that, the proposed method achieves better performance and higher robustness on domain text segmentation compared with the state-of-the-art approach, especially on the recall of OOV (out-of-vocabulary).

Key words: Chinese Word Segmentation；Domain Adaptation; Neural Network

1 引言

长久以来，分词一直是中文信息处理中的经典研究问题。在目前主流的方法中，分词被视为一种序列标注任务[1]。一些概率图模型陆续被成功应用在分词建模问题中，并结合研究者精心设计的各类特征[2-5]，获得了良好的效果，尤其是条件随机场（Conditional Random Fields，CRF）[6]已成为当前分词模型的主流。近年来，随着深度学习方法在图像语音等应用场景中大获成功，越来越多的学者开始将深度学习引入自然语言处理领域。已有一些学者尝试在中文分词问题上引入深度学习方法，并取得了一定的进展[7-12]。基于深度学习的中文分词方法通过直接从数据中端到端得训练，在无手工设计特征的情况下也可以取得较好的

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61472436，61532001）；

作者简介：吴佳林（1993——），男，硕士研究生，主要研究方向为自然语言处理和序列标注；唐晋韬（1981——），男，讲师，主要研究方向为信息抽取和自然语言处理；李莎莎（1982——），女，讲师，主要研究方向为自然语言处理和社交网络分析；王挺（1970——），男，教授，主要研究方向为自然语言处理和语义 web。

分词效果。

以上这些分词方法在各自的实验设置中均取得了较高的准确率（95%以上），但当训练集和测试集处在不同领域[13]或分词标准下时，准确率会显著下降。此外，如果需要针对特定领域获得有效的分词模型，以上方法均需要大量的人工标注语料，极大的限制了上述方法的实际应用。

学术界已有一些面向领域适应的分词工作。比如，有学者通过向分词模型中引入多种领域特征[14,15]来提高领域适应性。Zhang[16]等人提出了一种基于类型监督的面向中文分词和词性标注任务的领域适应联合训练框架。Liu[17]等人通过扩展 CRF 模型的目标函数，使其可以同时利用全标注和部分标注的领域语料，降低了领域语料的标注要求和构建成本。

然而，上述面向领域适应的分词方法均需重新训练分词模型，依赖于源领域的训练语料，同时也与分词模型深度耦合。事实上，在进行领域应用时，用来训练源领域分词器的语料通常很难得到。此外，这些方法中对领域资源的利用方式受限于分词模型，比如在基于 CRF 的分词模型中，领域知识只能被编码成特征函数或是适应于模型目标函数的特定形式，更增大了领域适应工作的难度。

Huang[18]等人提出了一种松耦合的方法，将领域分词过程分解为两个步骤：通用分词过程和领域纠正过程。该方法不与分词模型耦合，克服了前述方法的缺点。然而，其领域纠正过程基于 CRF 模型实现，领域适应性能严重依赖于特征设计的质量，因此缺乏灵活性。

本文中，我们采用了 Huang[18]等人描述的领域适应两阶段框架，并在其中引入基于神经网络模型的纠正器（Neural Network based Corrector, NNC），通过纠正通用分词器的输出结果达到适应目标领域的目的。NNC 仅需小规模的领域标注语料，无需进行特征设计就能自动学习针对特定领域的纠正模式。由于不与分词模型耦合，我们的方法具有较好的灵活性，同时利用神经网络的特征学习能力，减少了对特征工程的依赖。除此之外，我们受到当前神经网络方法的启发，利用预训练初始化技术进一步提升了 NNC 的纠正性能。我们在两个不同领域的数据集上结合多个不同分词性能和分词标准的通用分词器，验证了本文方法的有效性。实验结果表明，我们的方法与同类方法的最好结果相比具有更高的性能，尤其在未登录词（Out-of-vocabulary, OOV）召回率方面提升显著（相对提高了 20%），并且在结合不同的通用分词器时展示出了更强的鲁棒性。

2 领域纠正器框架

我们的工作基于 Huang[18]等人描述的用于领域适应中文分词框架。本文中，我们将其称为领域纠正器框架（Domain Corrector Framework, DCF）。

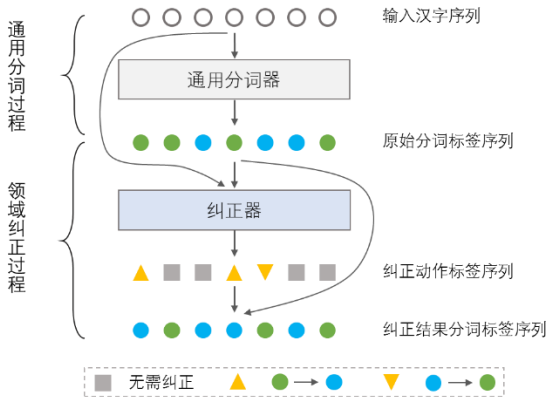

图 1：领域纠正器框架工作流程

领域纠正器框架由两部分组成：通用分词过程和领域纠正过程。图 1 展示了领域纠正框架的工作流程。其中，通用分词过程仅仅只对领域文本进行初步的分词，可由任意一种分词

器来完成。在获得初步分词结果后，领域纠正过程通过由统计模型实现的“纠正器”预测出针对通用分词结果的纠正动作序列，结合预先定义的纠正逻辑逐字修改分词结果，最终得到符合特定领域的分词输出。领域纠正器框架不依赖于构建通用分词器时使用的语料资源和建模方法，大大降低了对训练资源的要求，同时增强了引入领域知识的灵活性。下面对框架中的两个过程进行形式化说明。

2．1 通用分词过程

给定一个字符输入序列 $x_c = c_1, c_2, \cdots, c_n$ ，经过通用分词过程输出原始的分词结果，并将其转换为分词标签序列 $y_S = l_{S1}, l_{S2}, \cdots, l_{Sn}$ ，其中 $l_S \in L_S$ ，用来表示字符在词中的位置类型。文献[18]中，$L_S = \{\mathrm{B}, \mathrm{N}\}$，其中 B 代表词首字，N 代表词中字或词尾字。例如分词结果“我/爱/北京/天安门”对应的分词标签序列为“$\mathrm{B}, \mathrm{B}, \mathrm{B}, \mathrm{N}, \mathrm{B}, \mathrm{N}, \mathrm{N}$”。

2．2 领域纠正过程

上一过程结束后，x_c 和 y_S 同时被送入本文中称之为“纠正器”的纠正模块，以预测纠正标签序列 $y_C = l_{C1}, l_{C2}, \cdots, l_{Cn}$ ，其中 $l_C \in L_C$ ，表示纠正动作的类型。文献[18]中，$L_C = \{\mathrm{U}, \mathrm{I}, \mathrm{D}\}$ ，其中 U 代表无需纠正，D 代表删除字前空格，I 代表插入字前空格。例如对句子“苏如微笑道”的通用分词结果为“苏如微/笑道”，正确分词结果为“苏如/微笑/道”，则期望的纠正序列为“$\mathrm{U}, \mathrm{U}, \mathrm{I}, \mathrm{D}, \mathrm{I}$”。Huang[18]等人利用 CRF 模型对纠正器进行建模，因此纠正标签序列 y_C 在输入为 x_c 和 y_S 时的条件概率为：

$$P(y_C|\mathbf{x}) = \frac{1}{Z}\prod_{t=1}^{n}\exp\left\{\phi(y_{Ct}, \mathbf{x}) + \psi(y_{Ct}, y_{Ct+1}, \mathbf{x})\right\} \tag{1}$$

其中，$\mathbf{x} = (x_c, y_S)$，$\phi(y_{Ct}, \mathbf{x})$ 为标签在 t 位置的一元势函数，$\psi(y_{Ct}, y_{Ct+1}, \mathbf{x})$ 为 $t, t-1$ 位置标签之间的二元势函数，Z 为概率归一化因子。在测试阶段，由维特比算法可以求得条件概率最大的纠正标签序列 y_C^* 。最后，原始分词标签序列 y_S 通过相应的纠正序列 y_C^* 被修改为 $y_{\widetilde{S}}$ 。训练阶段，由已标注的目标分词序列 $\hat{y}_S$ 可以很容易推出相应的纠正序列 $\hat{y}_C$ ，再将 $\hat{y}_C$ 作为模型的目标输出，结合字符和分词标签这两种输入序列进行训练。

3 神经网络纠正器

从前一节的介绍中可以看出，纠正器是分词器进行领域适应的关键。然而文献[18]中提出的纠正器基于 CRF 模型，其特征模板除了要考虑字符上下文的不同模式，还需加入对分词标签信息的表达。如果要充分挖掘字符与分词标签之间的交互模式，还需要设计字符与分词标签的联合特征，更加大了特征设计的复杂性和难度。为了减少对特征工程的依赖，我们利用神经网络模型来自动学习源域分词和目标域分词之间的纠正模式。由于纠正动作大都只发生在分词规范不一致或未登录词的位置，多数位置的汉字无需纠正，因此纠正动作标签较为稀疏。与直接从小规模的领域数据中学习分词器相比，纠正器需要的模型容量相对较小。即使领域数据只有一个句子，过拟合带来的误纠正现象也只会发生在句中出现的汉字，并不会影响到大多数无需分词的位置。而如果学习分词器的模型只在一句上训练，那么该句之外的其他句子基本都会被分错。我们认为，虽然神经网络模型对训练数据量要求较高，但其在学习稀疏的纠正模式场景中是可用的，并且通过减少模型参数以及结合典型的正则方法可以进一步控制模型在小规模数据上的过拟合问题。为了能够有效捕捉长距离的纠正模式，同时约束输出标签之间的依赖关系，我们采用文献[19]中提出的双向 LSTM（Long Short-Term Memory）结合 CRF 的神经网络作为纠正器的基本模型。本文中，我们将 LSTM 单元替换为 GRU[20]（Gated Recurrent Unit），以减少纠正器的模型容量，进而在加快训练速度的同时减少模型过拟合的可能性。

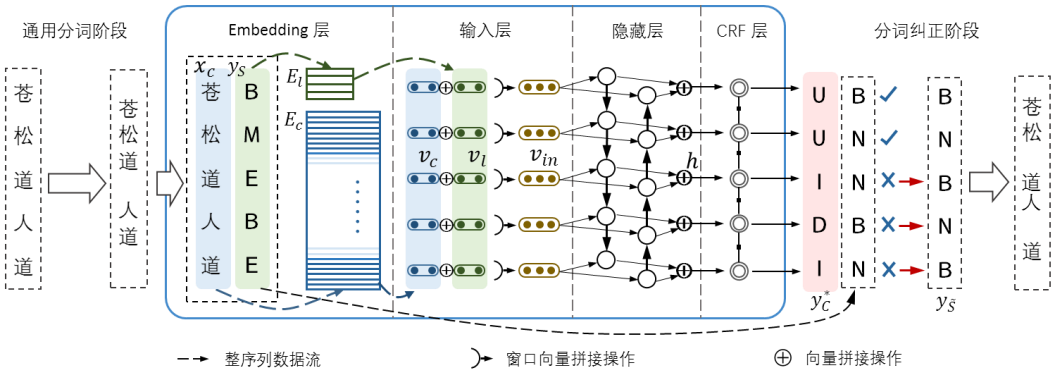

图 2：神经网络纠正器及其工作流程示例

如图 2 蓝色圆角方框中所示，神经网络纠正器的结构可分为 4 层：Embedding 层、输入层、隐藏层、CRF 层。

3．1 Embedding 层

该层由两种不同的 Embedding 组成:汉字 Embedding 和分词标签 Embedding,分别记为 $E_c \in \mathbb{R}^{|\mathrm{D}|\times d_c}$ 以及 $E_l \in \mathbb{R}^{|L_S|\times d_l}$，其中 $|\mathrm{D}|$ 表示汉字字典的大小。一个汉字字符 c_t 或一个分词标签 l_{St} 可分别通过 E_c 和 E_l 映射为一个维度为 d_c 以及 d_l 的数值向量 v_{ct} 和 v_{lt}。为了让模型得到更丰富的分词信息，我们没有跟从文献[18]中 2 词位标签集的设置，而是采用了具有更多样词位信息的 4 词位标签集 $L_S = \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}\}$，其中的标签可分别表示词首字（B）、词中字（M）、词尾字（E）以及单字词（S）。此外，该层可由预训练的 Embedding 来初始化。本文工作中，我们仅利用训练集语料生成汉字字符的 Embedding，并在后续实验中分别对比了有无预训练初始化情况下的模型性能。

3．2 输入层

在该层中，我们首先在每一序列位置 t 将 v_{ct} 和 v_{lt} 拼接为向量 v_{clt}，再将上下文窗口 $[t-k_1, t+k_2]$ 中的 k_1+k_2+1 个向量按顺序拼接为最终的输入向量v_{int}。即：

$$v_{clt} = [v_{ct}, v_{lt}] \tag{2}$$

$$v_{int} = [v_{clt-k_1}, v_{clt-k_1+1}, \cdots, v_{clt-k_2-1}, v_{clt-k_2}] \tag{3}$$

3．3 隐藏层

为了生成能够抽象上下文信息的隐藏表征，我们按照文献[19]中的描述，构建了双向循环神经网络（Bidirectional Recurrent Neural Network，Bi-RNN）。本文中，我们鉴于最终的实验效果和计算开销的考虑，将原始模型中 LSTM 单元替换为 GRU。GRU 单元和 LSTM 单元类似，都能有效得对序列中的长距离依赖关系进行建模。

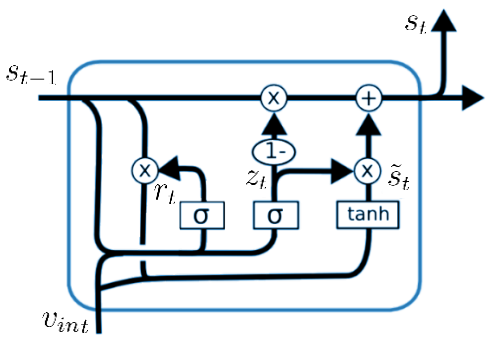

图 3：GRU 单元结构示意图

如图 3 所示，GRU 单元由重置门（reset gate）和更新门（update gate）来控制信息流动，相比 LSTM 单元具有更简单紧凑的结构和更少的待训参数，因此在数据规模较小的情况下

不易发生过拟合问题。通过双向 GRU 网络，我们可以从输入向量序列 v_{in} 得到上下文信息隐藏表征向量序列 h 。后文将省略所有线性变换中的偏执项以简化表达，具体计算公式为：

$$z_t = \sigma(W_z \cdot [s_{t-1}, v_{int}]) \tag{4}$$

$$r_t = \sigma(W_r \cdot [s_{t-1}, v_{int}]) \tag{5}$$

$$\tilde{s}_t = \tanh(W \cdot [r_t * s_{t-1}, v_{int}]) \tag{6}$$

$$s_t = (1 - z_t) * s_{t-1} + z_t * \tilde{s}_t = \mathrm{GRU}(s_{t-1}, v_{int}, W_z, W_r, W) \tag{7}$$

$$\overrightarrow{s_t} = \mathrm{GRU}(\overrightarrow{s}_{t-1}, v_{int}, \overrightarrow{W_z}, \overrightarrow{W_r}, \overrightarrow{W}) \tag{8}$$

$$\overleftarrow{s_t} = \mathrm{GRU}(\overleftarrow{s}_{t-1}, v_{int}, \overleftarrow{W_z}, \overleftarrow{W_r}, \overleftarrow{W}) \tag{9}$$

$$h_t = [\overrightarrow{s_t}, \overleftarrow{s_t}] \tag{10}$$

其中，符号 * 代表向量之间逐元素相乘；σ 为 sigmod 函数，可将任意实数映射到 0 到 1 之间的门控值；$\overrightarrow{s_t}$ 和 $\overleftarrow{s_t}$ 分别表示前向和后向 GRU 在 t 位置的输出状态向量。

3．4 CRF 层

与文献[19]相同，我们引入了一个线性链条件随机场模型作为网络的最后一层。隐藏表征向量序列 h 经过一个线性映射后，就得到由神经网络生成的一元势函数 $\phi_{nn}(y_{Ct}, \mathbf{x})$ 。在这里，二元势函数由一个可训练的矩阵 $A \in \mathbb{R}^{|L_C| \times |L_C|}$ 来表达，其中矩阵中的某一元素 $A_{i,j}$ 表示从纠正标签 i 转移到标签 j 的分值。因此，对于一个给定的纠正序列 y_C ，其由输入决定的条件概率可由下式计算：

$$s(\mathbf{x}, y_C) = \sum_{t=1}^{n} W_{\mathrm{output}}[y_{Ct}, :] \cdot h_t + A_{y_{Ct}, y_{Ct+1}} = \sum_{t=1}^{n} \phi_{nn}(y_{Ct}, \mathbf{x}) + A_{y_{Ct}, y_{Ct+1}} \tag{11}$$

$$P(y_C | \mathbf{x}; \theta) = \frac{1}{Z} \prod_{t=1}^{n} \exp\{\phi_{nn}(y_{Ct}, \mathbf{x}) + A_{y_{Ct}, y_{Ct+1}}\} \tag{12}$$

$$= \frac{e^{s(\mathbf{x}, y_C)}}{\sum_{\tilde{y} \in \mathbf{Y}} e^{s(\mathbf{x}, \tilde{y})}} \tag{13}$$

式（11）中，$W_{\mathrm{output}}[y_{Ct}, :]$ 表示矩阵 W_{output} 中标签 y_{Ct} 对应的行向量。网络以最小化正确序列 $\hat{y}_C$ 的负对数似然为目标，接受输入 $\mathbf{x}$ 进行端到端得优化训练。整个网络的损失函数（忽略正则化项）表达式为：

$$\mathcal{L}(\mathbf{x}, y_C; \theta) = \log(\sum_{\tilde{y} \in \mathbf{Y}} e^{s(\mathbf{x}, \tilde{y})}) - s(\mathbf{x}, y_C) \tag{14}$$

测试阶段，由维特比算法可求得概率最大的纠正标签序列 y_C^* 。

如图2所示，在纠正阶段，我们首先将4词位标签（B, M, E, S）转换为2词位标签（B,N），随后根据相应的纠正标签（U,I,D）进行纠正操作。其中，U 表示无需修改，I 表示将 N 纠正为 B ，D 表示将 B 纠正为 N 。最后，由纠正后的 2 词位标签得到最终的分词结果。

4 实验

为了验证领域纠正框架的有效性，我们将预先训练好的通用分词器在选取的两个领域数据集上进行了分词性能测试，同时测试了结合纠正器后的分词性能，预期后者较前者有显著提高。为了验证本文提出的神经网络纠正器（NNC）优于文献[18]中的 CRF 纠正器，我们在不同分词标准和分词性能的通用分词器下分别在两个领域数据集上测试了两种纠正器的分词性能。实验设置以及实验结果如下：

4．1 实验设置

表 1：领域数据集统计详情

	领域训练集 + 开发集		测试集					
	句子数	词语数	句子数	词语数	OOV 比例 %			
					MSR + 领域训练集	PKU + 领域训练集	MSR	PKU
ZX	2209	67648	1214	34336	2.9	3.5	13.8	15.6
EF	2318	60361	1000	27104	2.8	2.7	5.7	5.0

4.1.1 数据集

我们在两个不同领域的全标注数据集上进行实验，每个数据集都被划分为训练集、开发集以及测试集。我们统计了测试集相对于其他两个数据集的 OOV 比例，用以评估在测试集上的泛化难度。在领域适应的实验中，用以计算 OOV 的词汇表由训练纠正器和通用分词器的数据中的词汇组成。本文使用由 SIGHAN2005 发布的两个简体中文分词语料库（PKU 和 MSR）作为通用分词器的训练集。为了评估在通用分词器训练语料上的泛化难度，我们也统计了 PKU 和 MSR 的测试集相对于其训练集的 OOV 比例。表 1 列出了关于两个领域数据集详细的统计信息，具体来说：

ZX ： 该数据集的文本取自一部著名的网络小说——《诛仙》，其中含有大量的非典型姓名、地点等等，如“田不易”、“鬼王宗”。这些命名方式不同于规范的通用语料，可以一定程度上代表武侠小说领域的特点。我们使用 Zhang[16]等人已标注好的训练集，将其随机划分为本文中使用的训练集（90%）和开发集（10%）。测试集与文献[16]中保持一致。

EF ： 我们在 CTB5 语料库中手工选取了 259 篇已标注的财经新闻组成了该数据集。该数据集中频繁出现中文数字词汇，股票名称等，而通用语料的数字大都由阿拉伯数字表达，因此该数据集具有明显的特征，可以一定程度代表财经领域。我们在 EF 数据文本中随机摘出 1000 句作为该领域的测试集，其余部分随机划分为该领域的训练集（90%）以及开发集（10%）。

4.1.2 评价指标

对于分词性能，我们使用 F1 值和 OOV 召回率来进行评价。F1 值由分词准确率和召回率的调和平均值得出。对于纠正器的鲁棒性，我们通过计算纠正器在对接不同类型的通用分词器时性能的标准差来进行估计，标准差越小则鲁棒性越好。

4.1.3 通用分词器

为了研究神经网络纠正器在对接不同质量的通用分词器时的性能表现，我们分别训练了一个较低分词性能的弱分词器和一个较高性能的强分词器。另外，通用分词器所遵循的不同分词标准对纠正器的影响也是我们希望研究的一个方面。比如，在 PKU 语料中，人名会被标注为姓和名（我们称之为细粒度分词），但在 MSR 语料中，姓和名作为一个整体而不会被切分开（我们称之为粗粒度分词）。因此，我们分别在 PKU 和 MSR 语料上训练了细粒度和粗粒度两个分词标准的分词器，每个分词器均有强弱两个版本。

表 2：通用分词器在源领域和两个目标领域的分词性能

分词器		源领域		ZX		EF	
		F1 %	OOV 召回率 %	F1 %	OOV 召回率 %	F1 %	OOV 召回率 %
粗粒度	弱性能	93.0	56.8	76.9	48.5	80.1	27.4
	强性能	96.2	69.7	79.3	49.4	81.5	29.3
细粒度	弱性能	90.8	51.2	76.0	51.6	87.7	37.7
	强性能	94.3	70.0	81.0	57.5	90.3	54.5
平均值		93.6	61.9	78.3	51.8	84.9	37.2

表 2 列出了所有通用分词器在源领域和目标领域上的分词性能。正如前文所述，当训练集与测试集不在同一领域时，通用分词方法的性能会显著下降。在 ZX 数据集上，通用分词器性能下降最为显著，F1 值百分比最高下降了 16.9，平均为 15.7。而在 EF 数据集上，性能损失有所减弱，F1 值最高下降 14.7，平均为 8.7。我们分析，ZX 数据集来源于仙侠小说，而训练通用分词器的源领域数据大部分来源于正式的新闻报刊文本，领域相似性较低；EF 数据集由财经新闻构成，与源领域较为相似。因此通用分词器在 ZX 数据集上的分词效果要低于 EF 数据集。另外，我们发现不同分词标准的分词器在 ZX 数据上的性能差距较小，而在 EF 数据上的差距较大。由于 EF 数据集的分词标准偏向细粒度，细粒度的强通用分词器在 EF 数据上的 F1 值只下降了 4 个百分点，而在粗粒度的强通用分词器下，F1 值百分比减小了 14.7，差距将近 4 倍。我们推测，领域相似性越高，分词标准差异对领域适应的影响越大。

在未登录词召回率方面，通用分词器在 ZX 数据集上平均下降了 10.1 个百分点，而在 EF 数据集平均下降了 24.7，这与 F1 值在两个数据集上的情况恰好相反。为了分析该现象，我们以粗粒度强分词器的分词结果为例，分别统计了两个数据集测试文本中各未登录词提及次数占未登录词总提及次数的比例，将其作为评价词语重复度的定量指标。如图 4 所示，我们选取了重复度（提及次数占比）最大的 20 个词语，同时也统计了各词语在通用分词器下的召回比例。由统计结果可以发现，由于 ZX 文本取自同一部小说，未登录词重复度高的词较多（比如主角姓名、门派、折线引号等），这些重复词被正确切分对未登录词召回率的贡献较大。而通用分词器对规范的人名、地名的分词准确率较高，因此 ZX 中多个重复度高的词语（例如“张小凡”、折线引号、“小周”等）召回比例大。EF 文本取自不同的财经新闻，大部分的未登录词为中文数字，其重复度较低，因此其被正确切分对召回率的贡献较小，使得 EF 的未登录词召回率反而低于 ZX。但是，未登录词在一定程度上反应了领域的特征，因此可以认为领域相似度越大，模型在未登录词上的泛化难度也就越小，其召回率的提升潜力更大。这一点也在后文实验中得到证实。

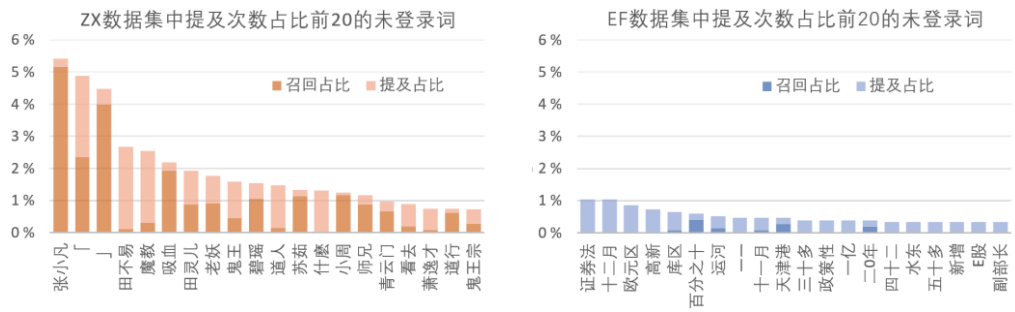

图 4：提及次数占比前 20 的未登录词及其召回比例

4.1.4 纠正器基准方法

基于纠正器的两阶段方法强调仅利用通用分词器的标注结果，而当前大多数领域适应方

法需要利用源领域语料，还可能依赖于源领域分词模型，与本文方法不具有可比性。因此，我们选取了文献[18]中基于 CRF 模型的纠正器进行比较，构造了与其相同的特征模板重新实现了 CRF 纠正器（简称为 CC）。该特征模板由三种类型的特征构成：“标签-标签”特征、“标签-词语”特征、“标签-分词”特征。

“标签-标签”特征表达了相邻输出标签之间的全局约束关系，类似于本文方法中用于建模二元势函数的标签转移矩阵。“标签-词语”特征表达了一定窗口范围内字符上下文与输出标签的局部约束关系。“标签-分词”特征表达了由通用分词器得到的原始分词标签与输出标签的局部约束关系，且原始分词标签不带上下文窗口。

关于 CRF 纠正器特征函数的详细说明见文献[18]，这里不予赘述。

4.1.5 超参数以及训练细节

通过实验，我们对多项超参数进行了调校，后文实验结果均基于该节参数设置。

字符向量和分词标签向量维数分别为 50 和 10，上下文窗口宽度 k_1 和 k_2 均取 1。隐藏层 GRU 单元状态维数为 50。我们在模型中使用了 2 种正则化方法，分别为代价函数中的 L2 正则项约束(权重为10^{-5}),以及添加在输入层与隐层之间的 Dropout 机制（比率为 0.2）。

模型训练方面，我们使用了 Adam[21]作为自适应学习率的梯度更新方法,初始学习率为 0.005，mini-batch 数为 20。同时，我们在采用了提前终止方法来预防过拟合，规定在验证集上，F1 值连续 5 次未超过历史最高值时停止训练，并输出最高 F1 值对应的模型。

实现方面，我们采用 CRF++[1]实现基准方法，同时利用 Tensorflow 作为神经网络模型的实现框架。另外，我们使用了 gensim Python 工具包中的 CBOW[22]词向量算法作为训练本文字向量的方法。

4．2 领域适应性实验结果

表 3：纠正器实验结果

	通用分词器		F1 %			OOV 召回率 %		
			CC	NNC	NNC+预训练	CC	NNC	NNC+预训练
ZX	粗粒度	弱性能	90.5	91.2	**91.6**	54.2	57.7	**61.6**
		强性能	91.8	91.9	**92.4**	58.4	62.4	**65.1**
	细粒度	弱性能	90.1	90.9	**91.4**	61.5	63.2	**65.5**
		强性能	91.4	91.0	**91.6**	64.8	65.4	**67.6**
	平均值（标准差）		91.0 (0.79)	91.3 (0.45)	**91.8 (0.44)**	59.7 (4.52)	62.2 (3.24)	**65.0 (2.47)**
EF	粗粒度	弱性能	94.2	**95.4**	**95.4**	65.5	77.4	**78.7**
		强性能	94.8	**95.8**	95.4	66.8	**78.9**	**78.9**
	细粒度	弱性能	94.6	95.0	**95.6**	70.3	74.4	**77.3**
		强性能	**95.7**	95.0	95.3	73.5	74.7	**76.5**
	平均值（标准差）		94.8 (0.63)	95.3 (0.38)	**95.4 (0.13)**	69.0 (3.61)	76.4 (2.17)	**77.9 (1.41)**

表 3 列出了领域分词性能和分词鲁棒性的评估结果。为了研究预训练字向量是否会带来进一步的性能提升，我们在所有神经网络纠正器的实验中都进行了测试。值得注意的是，实验中字向量的训练仅利用了纠正器的训练语料，并不包含任何训练集之外的资源，因此本文实验属于封闭测试的范畴。

与表 2 中展示的通用分词器性能相比，纠正器显著提高了在目标领域分词结果的 F1 值。在仅使用一个小规模标注集的条件下，即使最差的纠正器在两个领域中也均超过通用分词器 10%以上。实验结果充分证明了领域纠正器框架的有效性。另外，纠正器在 EF 数据集上的 OOV 召回率平均高于 ZX 数据集 10%，这也印证了 4.1.3 结尾的分析结论。

[1] https://taku910.github.io/crfpp/

在纠正器的性能方面，实验结果显示，在大多数情况下神经网络纠正器都超越了 CRF 纠正器。另外，经过预训练字向量初始化的神经网络纠正器的性能有了进一步的提高，特别是在 OOV 召回率方面提升显著。在预训练初始化后，相比 CRF 纠正器，神经网络纠正器的 OOV 召回率分别在 ZX 和 EF 领域上平均相对提高了 8.7%和 12.8%，并且在最好的情况下提高了 20%。这充分证明了神经网络纠正器相比 CRF 纠正器具有更强的泛化能力。

在鲁棒性评估方面，我们计算了同领域中在不同通用分词器下纠正器性能（F1 值以及 OOV 召回率）的标准差，发现神经网络纠正器与 CRF 纠正器相比具有更低的标准差，这意味着神经网络纠正器的鲁棒性更强。经过预训练的字向量初始化后，其鲁棒性得到了进一步的提升。

除了在完整训练集上进行实验，本文还研究了纠正器模型在不同训练集规模下的性能表现。我们以纠正粗粒度分词器在 ZX 数据集上的分词为例，将待实验的纠正器模型分别在不同比例规模的训练集上训练，比例分别取 1%、2%、5%、10%、20%、30%、40%、60%、80%以及 100%。图 5 展示了各纠正器分词性能（F1 值和 OOV 召回率）随训练集比例增加的变化曲线。横坐标为训练集比例，纵坐标为性能分值。其中“NNC+Pre”表示经过预训练字向量初始化的神经网络纠正器。横轴 0 点处的分值对应表 2 中不加纠正器的原始分词性能。

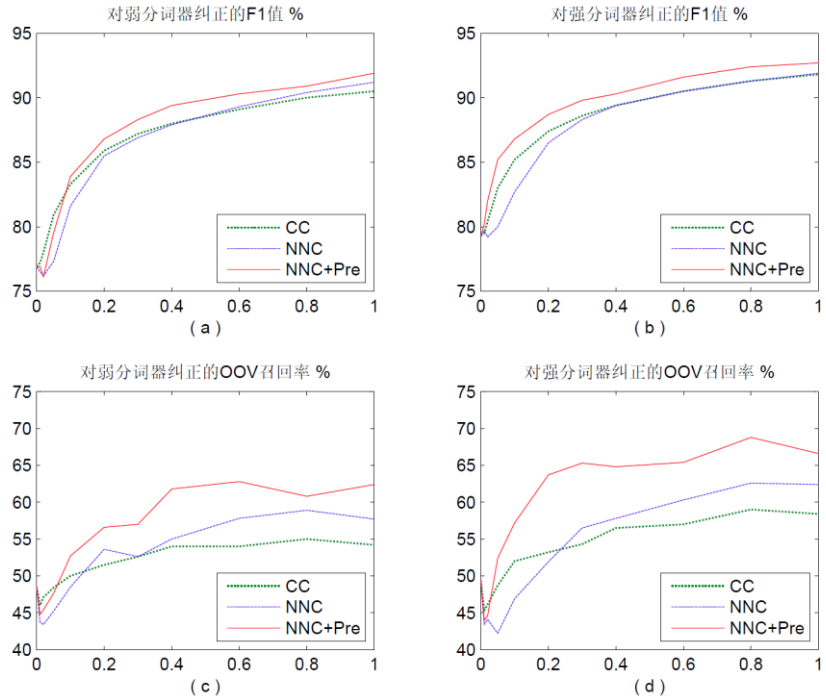

图 5：纠正器在不同比例 ZX 训练集下的分词性能（源领域为粗粒度分词）

可以看到，各纠正器的分词性能总体上与训练集规模成正相关。另外，NNC 在训练集规模超过 50%的情况下，两项性能均超过或不低于 CC。但当训练集规模较小时，无论是 F1 值还是 OOV 召回率，NNC 均低于 CC，且在训练集刚开始增加时性能突降，形成深沟型曲线。我们分析，NNC 基于神经网络模型，其参数较多，且没有 CRF 中特征模板的先验指导，在较小规模数据（不多于 100 句）中训练时无法有效形成对数据特征的抽象能力，容易发生过拟合问题。随着训练数据的增加，NNC 开始展现其优势。由于回避了特征设计问题，NNC 可以从原始数据中自动拟合纠正规律，结合双向 GRU 网络对序列数据的抽象建模能力，充分挖掘了整句中所有原始分词标签和汉字表征之间的交互关系，因此体现出了更强的泛化能力。在引入预训练字向量后，NNC 的性能得到了显著提高。特别是在训练集规模较小时，分词性能的突降现象得到了有效缓解，在规模比例不大于 10%时分词性能就超过了 CC。这说明，即使在小规模数据上训练的字向量，也可以提高 NNC 的分词性能。

5 结论与未来工作

本文提出了一种基于神经网络的中文分词方法，以提高分词系统向新领域迁移的适应性和灵活性。由于不与分词模型耦合，我们的方法具有较好的灵活性，同时利用神经网络的自动抽象能力，减少了对特征工程的依赖。实验结果表明，神经网络纠正器显著提高了在新领域中的分词性能，尤其在 OOV 召回率方面提升显著（最高相对提升了 20%）。除此之外，该方法相比已有的纠正器方法具有更高的鲁棒性。在未来的研究中，我们将尝试扩展纠正器模型，使其能够充分利用领域无标注语料和领域词典，减少对全标注语料的依赖，进一步提高方法的实用性。

参考文献：

[1] Xue N, Shen L. Chinese word segmentation as lmr tagging[C]. Sighan Workshop on Chinese Language Processing, 2003: 176-179.

[2] Tseng H, Chang P, Andrew G, et al. A conditional random field word segmenter[J]. Foundations of Science. 2005, 168–171.

[3] Zhao H, Huang C N, Li M, et al. A unified character-based tagging framework for Chinese word segmentation[J]. Acm Transactions on Asian Language Information Processing. 2010, 9(2):1–32.

[4] Sun W, Xu J. Enhancing chinese word segmentation using unlabeled data[C]. Conference on Empirical Methods in Natural Language Processing, 2011: 970-979.

[5] Sun X, Wang H, Li W. Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection[C]. ACL, 2013: 253-262.

[6] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[M]: volume 3. [S.l.]: [s.n.], 2001: 282–289.

[7] Zheng X, Chen H, Xu T. Deep learning for chinese word segmentation and pos tagging.[C]. EMNLP, 2013: 647–657.

[8] Pei W, Ge T, Chang B. Max-margin tensor neural network for chinese word segmentation[C]. Meeting of the Association for Computational Linguistics, 2014: 293-303.

[9] Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for chinese word segmentation.[C]. ACL (1), 2015: 1744–1753.

[10] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for chinese word segmentation[C]. Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.

[11] Cai D, Zhao H. Neural word segmentation learning for chinese[C]. Meeting of the Association for Computational Linguistics, 2016: 409-420.

[12] Yao Y, Huang Z. Bi-directional lstm recurrent neural network for chinese word segmentation[C]// Springer. International Conference on Neural Information Processing: Springer, 2016: 345–353.

[13] Liu Y, Zhang Y. Unsupervised domain adaptation for joint segmentation and pos-tagging.[C]. COLING (Posters), 2012: 745–754.

[14] Chang B, Han D. Enhancing domain portability of chinese segmentation model using chi-square statistics and bootstrapping[C]. Conference on Empirical Methods in Natural Language Processing, 2010: 789-798.

[15] 韩冬煦, 常宝宝. 中文分词模型的领域适应性方法[J]. 计算机学报, 2015, 38(2):272-281.

[16] Zhang M, Zhang Y, Che W, et al. Type-supervised domain adaptation for joint segmentation and pos-tagging.[C]. EACL, 2014: 588–597.

[17] Liu Y, Zhang Y, Che W, et al. Domain adaptation for crf-based chinese word segmentation using free annotations.[C]. EMNLP, 2014: 864–874.

[18] Huang L, Du Y, Chen G. Geosegmenter: A statistically learned chinese word segmenter for the geoscience domain[J]. Computers & Geosciences. 2015, 76:11–17.

[19] Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging[J]. arXiv preprint arXiv:1508.01991. 2015.

[20] Cho K, VanMerrienboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259. 2014.

[21] Kingma D, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980. 2014.

[22] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems, 2013: 3111–3119.

吴佳林（1993——），男，硕士研究生，主要研究方向为自然语言处理和序列标注。
Email:wujialin11@nudt.edu.cn

唐晋韬（1981——），男，讲师，主要研究方向为信息抽取和自然语言处理。
Email:tangjintao@nudt.edu.cn

李莎莎（1982——），女，讲师，主要研究方向为自然语言处理和社交网络分析。
Email:shashali@nudt.edu.cn

王挺（1970——），男，教授，主要研究方向为自然语言处理和语义 web。
Email:tingwang@nudt.edu.cn