基于感知器的中文分词增量训练方法研究*

韩冰[1]，刘一佳[1]，车万翔[1]，刘挺[1]

（1.哈尔滨工业大学计算机学院社会计算与信息检索研究中心 哈尔滨 150001）

摘 要：文本提出了一种基于感知器的中文分词增量训练方法。该方法可在训练好的模型基础上添加目标领域标注数据继续训练，解决了大规模切分数据难于共享，源领域与目标领域数据混合需要重新训练等问题。实验表明，增量训练可以有效提升领域适应性，达到与传统数据混合相类似的效果。同时本文方法模型占用空间小，训练时间快，可以快速训练获得目标领域的模型。

关键词：中文分词；领域适应；增量训练

中图分类号：TP391　　**文献标识码：**A

An Incremental-styled Learning Scheme for Perceptron based Chinese Word Segmentation

Bing Han[1], Yijia Liu[1], Wanxiang Che[1], Ting Liu[1]

(1.Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001)

Abstract: In this paper, we propose an incremental-styled learning scheme in perceptron based Chinese word segmentation. Our method can perform continuous training over a fine tuned source domain model. Such scheme allows delivering model without annotated data and without re-training on these data. Experimental results shows the scheme we proposed can significantly improve adaptation performance on Chinese word segmentation and achieve comparable performance with traditional method. At the same time, our method can significantly reduce the resulted model size and obtain segmentation model with less time consumption.

Key words: Chinese Word Segmentation; Domain Adaptation; Incremental Learning

1 引言

词是汉语中的最小语义单元。由于，汉语以字为基本的书写单位，词与词之间没有明显的分割标记，中文分词成为中文信息处理的基础与关键，在信息检索、文本挖掘等任务中广泛使用。近年来，基于统计的中文分词方法在新闻领域取得了很好的性能[1-4]。但随着互联网、社交媒体与移动平台的迅猛发展，当前中文分词方法处理的数据不单局限于新闻领域。不断增长的开放领域数据对中文分词方法提出了新的挑战。前人研究[5-7]表明，在新闻领域训练的中文分词模型切换到诸如论坛、微博、小说等领域时，性能往往严重下降。

前人工作[6]将这种训练与测试领域不一致致使模型性能下降的问题归纳为领域适应问题。在使用新闻领域训练的分词模型处理开放领域时，新闻领域为源领域，开放领域为目标领域。出现这种问题主要有两点原因：一是不同领域数据文体不一致，例如小说与新闻使用不同的语言风格；二是不同领域间领域词典不一致，如金融领域经常使用“做空”“配资”等新闻领域不常用的词汇。Liu 和 Zhang[6] 通过在分词词性标注联合模型上加入聚类特征的方式捕捉源领域与目标领域的相似性，以解决文体差异过大问题。Zhang 等[5]将目标领域词典融入模型，避免了源领域与目标领域词典差异过大。Liu 等[7]提出了一种利用网络文本中自然存在的分词边界的方法，在基于条件随机场（CRF）模型的分词系统上提高了领

* **收稿日期：**　　**定稿日期：**

基金项目： 2014CB340503

域适应性。

上述研究表明，使用目标领域切分数据训练模型以解决领域适应问题的方法是一种高精度的方法。同时，在源领域切分数据的基础上加入目标领域数据这类混合训练数据的方法可以进一步提高切分中文分词准确率[5,7,8]。然而，多方面因素限制了这一类方法的适用性。其一，大规模切分数据往往很难公开，使得混合训练数据的方法难以应用于实际场景；其二，针对每个目标领域的混合数据都需要在包含源领域的大规模数据上重新训练模型，使得这种方法很难快速获得模型并部署。

针对上述问题，本文提出一种了基于感知器的中文分词增量训练方法。本文方法通过在已有模型的基础上继续训练，可以在不需要源领域切分数据的情况下，利用少量目标领域标注数据获得与混合模型相近的性能。同时本文针对增量训练提出了一种优化的实现方法，显著降低了训练代价。本文分词器将在 https://github.com/HIT-SCIR/ltp 开源。

2 问题描述

文本主要解决应用场景下的中文分词领域适应问题（示意图见图 1）。本文假设源领域数据在训练领域适应模型时对用户不可见，但源领域模型可见。本文同时假设用户有少量目标领域标注数据。最后，本文假设源领域模型同时服务于多个目标领域。

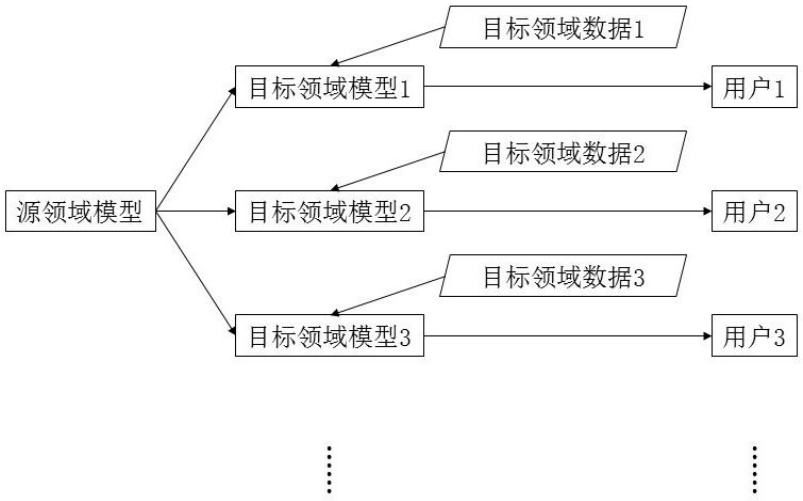


图 1 多领域应用场景示意图

针对以上问题描述，本文训练算法应具有下述特点：

- 不更改源领域模型
- 目标领域模型与混合数据训练的模型性能相近
- 目标领域模型精简

3 基于感知器的中文分词

本文参照前人工作[3,9]，将中文分词建模为基于字的序列标注问题。模型给句子中的每个字标注一个表示词边界的标记。本文采用了{B、I、E、S}四种标记，其中 B 代表词语的开始，I 代表词语的中间，E 代表词语的结尾，S 代表单个字词语。以“总理李克强调研上海外高桥”为例，标注结果如图 2 所示。

总	理	李	克	强	调	研	上	海	外	高	桥
B	E	B	I	E	B	E	B	E	B	I	E

图 2 分词序列标注示例

本文采用结构化感知器模型(Structured Perceptron[10])训练。为了防止模型过拟合，采用

平均感知器算法对训练过程中的参数求平均。基于感知器的中文分词训练算法如算 1 所示。

算法 1 平均感知器模型训练算法

输入：$D = \{(x,y)\}_N$
$w \leftarrow 0$
for $t = 1 \ldots T$ do
  for $(x_i, y_i) \in D$ do
    $z = \text{argmax}_{y' \in GEN(x_i)}(\phi(x_i, y') \cdot w)$
    if $z \neq y$ then
      $w \leftarrow w + \phi(x_i, y_i) - \phi(x_i, z)$
    end if
  end for
end for
$\overline{w} = \frac{1}{NT} \sum_{n=1..N,t=1..T} w^{n,t}$
return $\overline{w}$

其中，N 表示训练样例个数，T 表示迭代训练次数，(x_i, y_i)表示第 i 个训练数据，$\phi(x,y)$表示特征函数，把输入映射为向量，$GEN(x_i)$表示在输入是x_i时所有可能的输出，w 表示参数向量，$\overline{w}$表示平均化的参数向量，$w^{n,t}$表示第 t 轮迭代第 n 个数据的参数向量。

4 平均感知器增量训练算法

为了解决重复训练，领域数据快速更迭等问题，本文在结构化感知器中文分词的基础上提出一种增量式训练算法。

4．1 算法

本文方法可以归纳为在已有的感知器分词模型基础之上继续训练。增量式训练算法包含两个阶段：第一阶段的训练算法与传统感知器算法相同，用数据集D_1训练得到模型w_1；第二阶段，用数据集D_2和模型w_1训练模型得到模型w_2（如算法 2 所示）。

算法 2 感知器模型增量训练算法

$w_1 \leftarrow$ perceptron-train(D_1)
$w_2 \leftarrow$ incremental-perceptron-train(w_1，D_2)

在实际应用情景中，D_1是相对丰富且不同于目标领域的标注数据，例如新闻领域数据；D_2是目标领域的相对较少的标注数据，如财经、小说等。第二阶段的训练算法，以模型w_1和目标领域数据D_2为输入。设D_1有N_1条数据，第一阶段迭代训练T_1次，第二阶段同理，$w^{n,t}$表示在第 t 轮更新第 n 个数据时的参数向量，则第二阶段的平均参数为

$$\overline{w} = \frac{1}{N_1T_1 + N_2T_2}\left(\sum_{n_1=1..N_1,t_1=1..T_1} w^{n,t} + \sum_{n_2=1..N_2,t_1=T_1..T_1+T_2} w^{n,t}\right)$$

4．2 增量训练收敛性的证明

Collins 等人[10]证明了结构化感知器算法的收敛性。本文提出了一种增量训练算法，需要回答“增量训练算法能否在D_2数据上有限步骤内收敛”，亦即证明其收敛性。由于增量训练采用第一阶段的模型参数做为初始，证明增量训练的收敛性的问题等价于证明感知器算法在初始权重$w^1 \neq 0$时的收敛性。本文沿用 Collins 等人[10]的证明方法，在这一段证明D_2线性可分的情况下收敛。

定义： $\overline{GEN(x_i)} = GEN(x_i) - y_i$为除去正确序列$y_i$之外，所有可能的序列。向量 U，满足$\|U\| = 1$且$\forall_i \forall_z \in \overline{GEN(x_i)}$，$U \cdot \phi(x_i, y_i) - U \cdot \phi(x_i, z) \geq \delta$。 w^k为第 k 次出错更新前的模型参数。

定理： 增量训练算法在D_2线性可分情况下收敛。

证明： 在增量训练的设置下，$w^1 \neq 0$，假设在分析第 i 个实例时第 k 次出错，令 z 为模型预测结果，其中$z = \text{argmax}_{y' \in GEN(x_i)} \phi(x_i, y') \cdot w$，算法将依照$w^{k+1} = w^k + \phi(x_i, y_i) - \phi(x_i, z)$更新参数。则$Uw^{k+1} = Uw^k + U(\phi(x_i, y_i) - \phi(x_i, z))$，令$\delta$为$\phi(x_i, y_i) - \phi(x_i, z)$的最小值，递推可得$Uw^{k+1} - Uw^1 \geq k\delta$，则$\|w^{k+1} - w^1\| \geq k^2\delta^2$，下界得证。

接下来，本文证明$\|w^{k+1} - w^1\|$的上界。$w^{k+1} - w^1 = w^k + (\phi(x_i, y_i) - \phi(x_i, z)) - w^1$，则$\|w^{k+1} - w^1\|^2 = \|w^k - w^1\|^2 + 2(w^k - w^1)(\phi(x_i, y_i) - \phi(x_i, z)) + ||\phi(x_i, y_i) - \phi(x_i, z)||^2$根据感知器更新准则，算法只有在出错，即$w^k \cdot \phi(x_i, y_i) \leq w^k \cdot \phi(x_i, z)$时更新参数。 因此$w^k(\phi(x_i, y_i) - \phi(x_i, z)) \leq 0$。所以，$\|w^{k+1} - w^1\|^2 \leq \|w^k - w^1\|^2 - 2w^{1\,T}(\phi(x_i, y_i) - \phi(x_i, z)) + \|\phi(x_i, y_i) - \phi(x_i, z)\|^2$。令$\|\phi(x_i, y_i) - \phi(x_i, z)\|^2 \leq R^2$，且$\mu$为$2w^{1\,T}(\phi(x_i, y_i) - \phi(x_i, z))$的最小值。则有$\|w^{k+1} - w^1\|^2 \leq \|w^k - w^1\|^2 - \mu + R^2$。递推可得$\|w^{k+1} - w^1\|^2 \leq -\mu k + kR^2$。所以$k \leq \frac{R^2 - \mu}{\delta}$。因为 k 小于等于一常量，可知增量训练算法可以在D_2上以有限步骤收敛。

同理可证线性不可分的情况下增量训练依旧收敛，限于篇幅限制该证明省略。

4．3 优化的增量训练实现方法

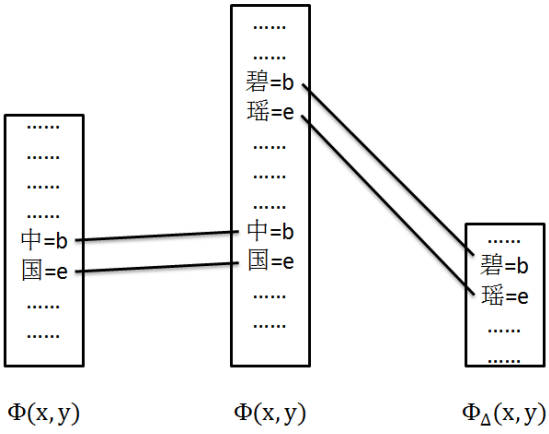


图 3 特征空间对比图

上述增量训练算法，第二阶段先复制创建一个与w_1一样的模型，并在此基础上增添训练语料D_2迭代更新参数，最终输出一个新的模型w_2。然而，在第二阶段仅更新了在D_2语料中出现的特征对应的参数，完全复制一份w_1在空间上是十分低效的。为此本文提出了一种更高效的实现方法。在第二阶段，创建一个新的模型w_Δ，该模型用来记录原始第二阶段训练的参数改变量，新模型w_Δ的工作依赖于w_1。原始领域特征空间、混合训练特空间与增量训练特空间如图 3 所示。由于第二阶段仅更新了D_2中出现的参数，因此增量模型w_Δ只需记录与D_2相关的参数，大大缩小了空间。优化后的增量训练第二阶段算法如算法 3 所示。

算法 3 优化增量训练第二阶段算法

输入：$D_2 = \{(x,y)\}_{N_2}, w_1$
for $t = T_1 + 1 \ldots T_2$ do
  for $(x_i, y_i) \in D_2$ do
    $z = \mathrm{argmax}_{y' \in GEN(x_i)}(\phi(x_i, y') \cdot w + \phi_\Delta(x_i, y') \cdot w_\Delta)$
    if $z \neq y$ then
      $w_\Delta \leftarrow w_\Delta + \phi_\Delta(x_i, y_i) - \phi_\Delta(x_i, z)$
    end if
  end for
end for
$\overline{w_\Delta} = \frac{1}{N_1 T_1 + N_2 T_2}\left(\sum_{n=1..N_1, t=1..T_1} w^{n,t} + \sum_{n=1..N_2, t=T_1..T_1+T_2.} w^{n,t}\right)$
return $\overline{w_\Delta}$

5 实验

5．1 实验设置

本文在 CTB5.0 和诛仙网络小说数据上进行试验。CTB5.0 数据划分参照前人工作[11]，用于训练第一阶段模型。诛仙小说数据划分参照 Zhang 等[5]，训练集用于训练第二阶段模型，测试集用于评价模型性能。为了模拟不同训练数据规模下算法的性能，随机选取 500 句诛仙训练数据作为小规模训练集，并用全部训练数据作为大规模数据。

在基于字的分词模型的特征方面，本文参考[12]的论文，并从一定程度上化简了其中的词典特征。本文的分词器使用的特征列表如表 1 所示。

表 1: 分词器使用的特征

编号	特征类别	特征模板
1	字的 n-gram 特征	c_i $(i = -2,-1,0,1,2)$
2		$c_i c_{i+1}$ $(i = -2,-1,0,1)$
3		$c_i c_{i+2}$ $(i = -1,0)$
4		$c_i c_{i+1} c_{i+2} (i = -1)$
5	字的重复信息特征	$\mathrm{dup}(c_i, c_{i+1}), (i = -1,0)$
		$\mathrm{dup}(c_i, c_i), (i = -1,0)$
6	字类别特征	$\mathrm{chartype}(c_0)$
7	词典信息特征	$\mathrm{match\_prefix}(c_0, D)$
		$\mathrm{match\_mid}(c_0, D)$
		$\mathrm{match\_suffix}(c_0, D)$

其中，下标 i 代表特征模板中的字与待标注字的相对位置。$\mathrm{dup}(x,y)$表示x,y是否为相同字，$\mathrm{chartype}(c)$表示c的字类型，字类型的定义包括字母（例如“A”），数字（例如“1”）以及标点（例如“，”）。本文使用的词典特征主要有三类，$\mathrm{match_prefix}(c_0, D)$表示以$c_0$为词首的句子片段在词典 D 中匹配的最长的词，$\mathrm{match_mid}(c_0, D)$表示以$c_0$为词中而$\mathrm{match_suffix}(c_0, D)$表示以$c_0$为词尾。本文使用的词典通过训练语料构造。构造方法是抽取训练语料中出现的频率大于等于 5 的词以及其词性构成词典。

5.2 增量训练实验

本文基线是系统使用 CTB5.0 训练数据训练的基于字的感知器中文分词模型，表 2 显示了基线模型的实验结果。本文分别在新闻（CTB5.0）和诛仙（ZX）测试集上评价基线模型性能。在与训练数据同源的新闻（CTB5.0）测试集上，基线模型的 F 值为 96.65%，而在诛

仙测试集上，F 值降到 86.55%。这说明单独由新闻领域数据训练的模型在诛仙数据集上存在领域适应问题。

表 2 基线分词模型实验结果

	CTB5.0	ZX
基线模型	96.65%	86.55%

为模拟不同规模目标领域的情况，本文分别采用随机选取的 500 句和 2400 句诛仙领域语料作为目标领域的训练数据。表 3 显示了不同方法利用两种规模训练数据训练的模型在诛仙测试集上的性能。第一行表示仅适用诛仙训练数据训练模型的情况下模型的性能。第二行表示使用新闻语料和诛仙语料混合训练获得的模型在诛仙领域上的性能。第三行表示使用本文提出的增量训练方法训练获得的模型的性能。

表 3 增量训练实验结果

	500	**2400**
仅适用诛仙训练数据	81.11%	94.09%
数据混合	**92.05%**	94.48%
增量训练	91.67%	**94.72%**
Zhang 等人（2014）[5]	94.61%	
Liu 等人（2014）[7]	90.63%	

通过对比表 3 第一行和第二行结果，可以得出结论，对于数据规模较小的领域，单独使用小规模数据并不能获得性能令人满意的模型。通过对比表 3 第二行和第三行结果，在小规模训练集上 F 值下降 0.39%，在大规模数据集上提升了 0.24%，结果表明二者性能相近。

本文也将实验结果与相同数据集上的前人工作进行了对比。本文提出的增量训练方法在 2400 句训练数据条件下，较 Zhang 等人[5]提出的当前准确率最好的模型获得了微小的提升。但由于 Zhang 等人使用的模型是分词词性标注联合模型，同时使用了词典以及自学习等策略。两者不具备直接考可比性。

5.3 实验分析

在关注增量训练准确率的同时，模型大小以及模型训练时间也是本文关注的一方面。本文经验性地比较了增量训练与传统混合训练的模型大小（表 4）。从表 4 可以看出，本文提出的优化实现方法可以显著减少模型大小。

表 4 模型大小

模型大小	500	2400
CTB5+ZX	30M	32M
增量训练 ZX	396K	4.4M

同时，本文比较了增量训练与混合数据方式训练的时间开销。在开发集上，本文将不同数据规模下增量训练的随时间的收敛曲线如图 4 所示。在小规模训练集上，增量训练相对于传统训练迅速达到最优结果。在大规模训练集上，二者趋于一致。

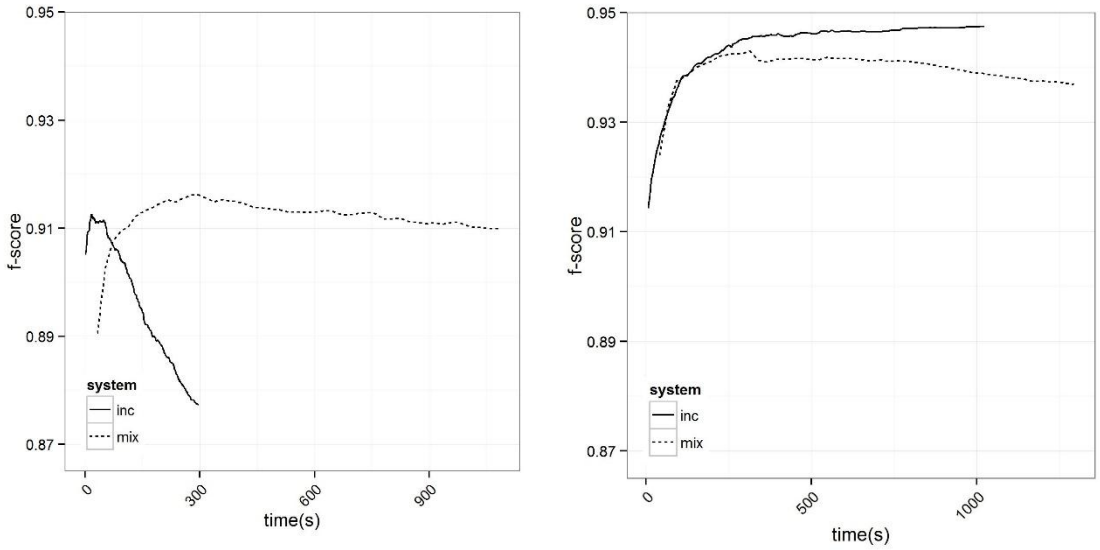


图 4 训练时间效率对比图左图为 500 句训练集的，右图为 2400 句训练集，图中横轴代表训练时间，单位为秒，纵轴为开发集上的 F 值。实线代表增量训练，虚线代表混合训练

上述实验表明，增量训练算法可以有效解决领域适应问题，通过在增量训练第二阶段添加目标领域语料，能有效提高在目标领域的性能。增量训练相对于传统混合训练方式，在准确性上基本持平，而在空间效率和时间效率上具有明显优势。

6 结论

针对领域适应问题，本文提出了一种增量训练算法来解决增加目标领域数据方法的限制。经过证明，增量训练算法可以在目标领域训练数据收敛。实验表明，通过在增量训练第二阶段添加目标领域训练语料，可以有效提升目标领域分词效果，并且增量训练算法模型占用的空间小，训练时间更快。

参考文献

[1] XUE N, SHEN L. Chinese word segmentation as LMR tagging[C]//Proceedings of the second SIGHAN workshop on Chinese language processing -. Morristown, NJ, USA: Association for Computational Linguistics, 2003, 17: 176－179.

[2] ZHANG Y, CLARK S. Chinese Segmentation with a Word-Based Perceptron Algorithm[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, Czech Republic: Association for Computational Linguistics, 2007: 840－847.

[3] SHI Y, WANG M. A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks[C]//Proceedings of IJCAI. 2007, 7: 1707－1712.

[4] SUN W. Word-based and Character-based Word Segmentation Models: Comparison and Combination[C]//Coling 2010: Posters. Beijing, China: Coling 2010 Organizing Committee, 2010(August): 1211－1219.

[5] ZHANG M, ZHANG Y, CHE W等. Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for

Computational Linguistics, 2014: 588－597.

[6] LIU Y, ZHANG Y. Unsupervised Domain Adaptation for Joint Segmentation and POS-Tagging[C]//Proceedings of COLING 2012: Posters. Mumbai, India: The COLING 2012 Organizing Committee, 2012, 2(December): 745－754.

[7] LIU Y, ZHANG Y, CHE W等. Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 864－874.

[8] LIU Y, ZHANG M, CHE W等. Micro blogs Oriented Word Segmentation System[C]//Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: Association for Computational Linguistics, 2012: 85－89.

[9] XUE N. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29－48.

[10] COLLINS M. Discriminative Training Methods for Hidden Markov Models: Theory and experiments with perceptron algorithms[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Morristown, NJ, USA: Association for Computational Linguistics, 2002, 10(July): 1－8.

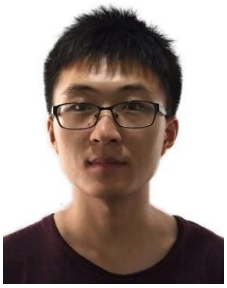
[11] SUN W, XU J. Enhancing Chinese word segmentation using unlabeled data[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011: 970－979.

[12] 张梅山, 邓知龙, 车万翔等. 统计与词典相结合的领域自适应中文分词[J]. 中文信息学报, 2010, 26(2): 8－12.

作者简介：韩冰（1990—），男，硕士生，主要研究方向为自然语言处理；刘一佳（1988—），男，博士生，主要研究方向为自然语言处理；车万翔（1980—），男，副教授，主要研究方向为自然语言处理；刘挺（1972—），男，教授，主要研究方向为自然语言处理，信息检索。

作者照片

韩冰



刘一佳 

车万翔 

通讯作者 刘挺 地址: 哈尔滨市南岗区教化街 29 号 6 楼 邮编: 150001 电子邮箱: tliu@ir.hit.edu.cn