文本数据增强方法剖析

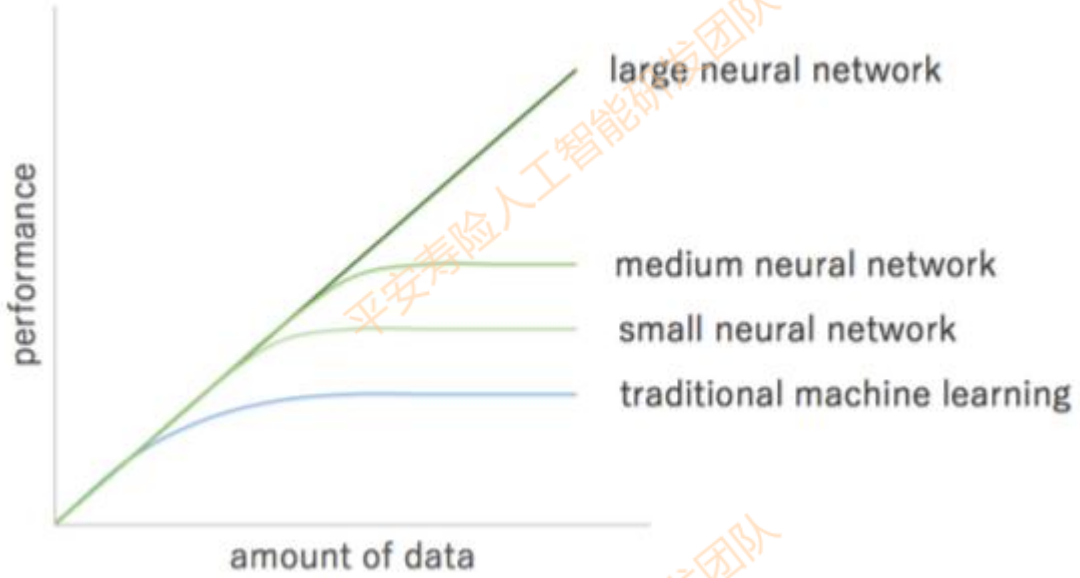
——如何快速处理文本数据不足及类别不均衡的问题

孙梦轩

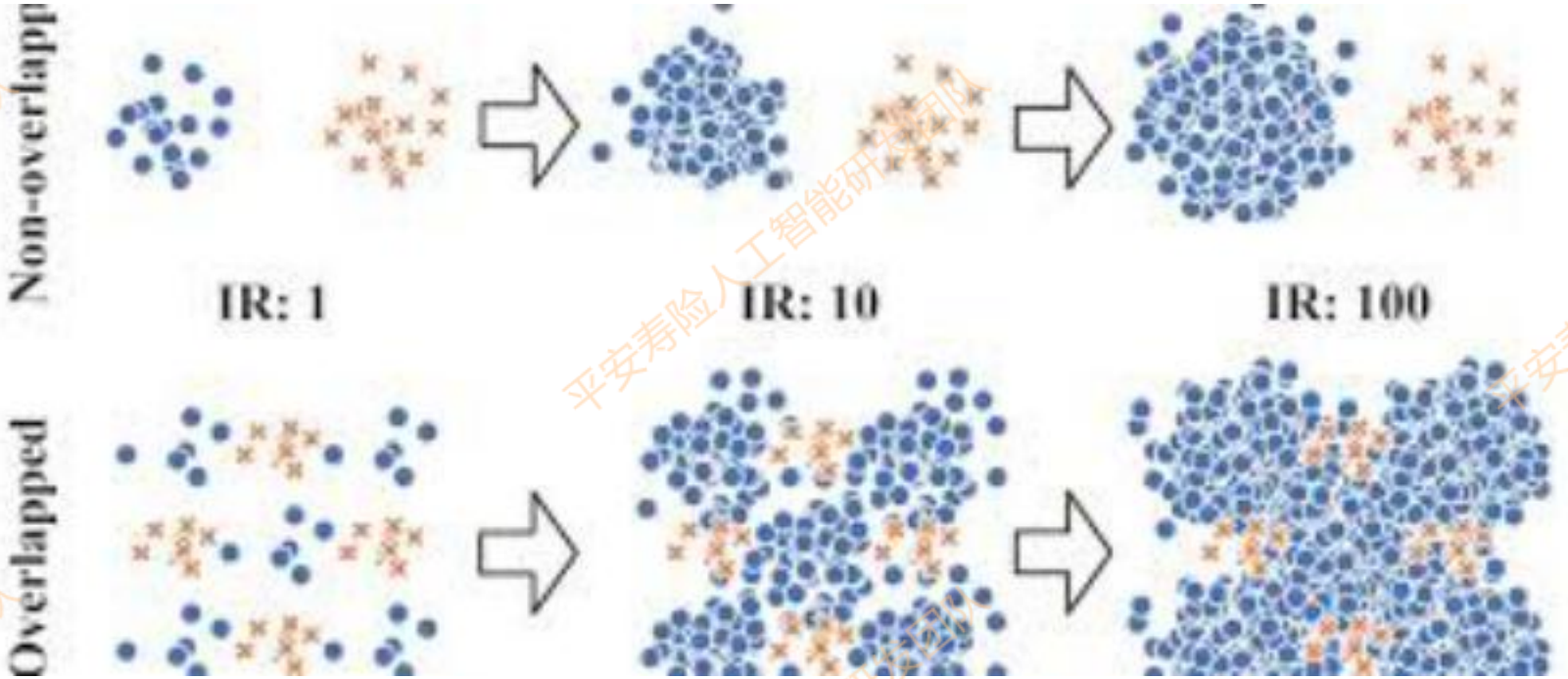
目录

一、数据增强的背景和应用场景

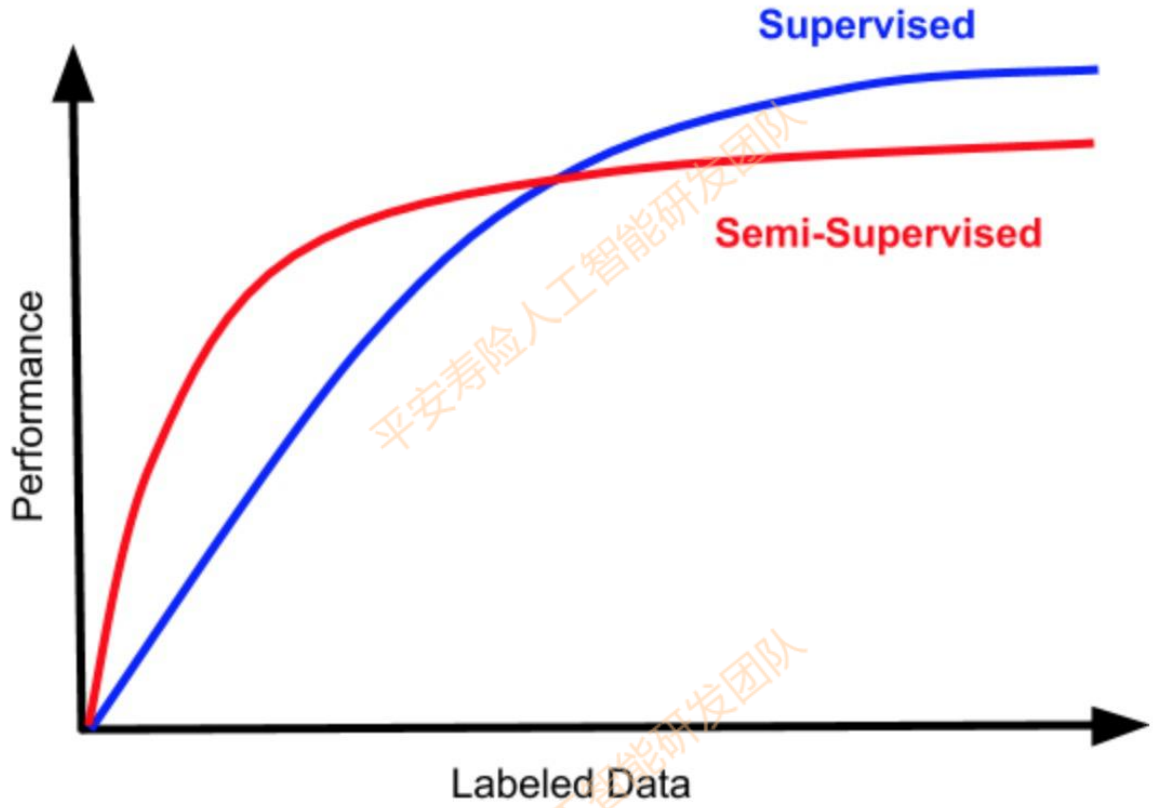
数据规模驱动机器学习发展



数据不平衡导致模型预测偏差



半监督相对有监督通常效果受限



二、传统文本数据增强的技术

2.1.1 Easy Data Augmentation for Text Classification Tasks

(1) 同义词替换（SR: Synonyms Replace）：

不考虑stopwords，在句子中随机抽取n个词，然后从同义词词典中随机抽取同义词，并进行替换。

Eg: “我非常喜欢这部电影” —> “我非常喜欢这个影片”，句子仍具有相同的含义，很有可能具有相同的标签。

(2) 随机插入(RI: Randomly Insert)：

不考虑stopwords，随机抽取一个词，然后在该词的同义词集合中随机选择一个，插入原句子中的随机位置。该过程可以重复n次。

Eg ：“我非常喜欢这部电影” —> “爱我非常喜欢这部电影影片”。

2.1.1 Easy Data Augmentation for Text Classification Tasks

(3) 随机交换(RS: Randomly Swap):

句子中，随机选择两个词，位置交换。该过程可以重复n次。

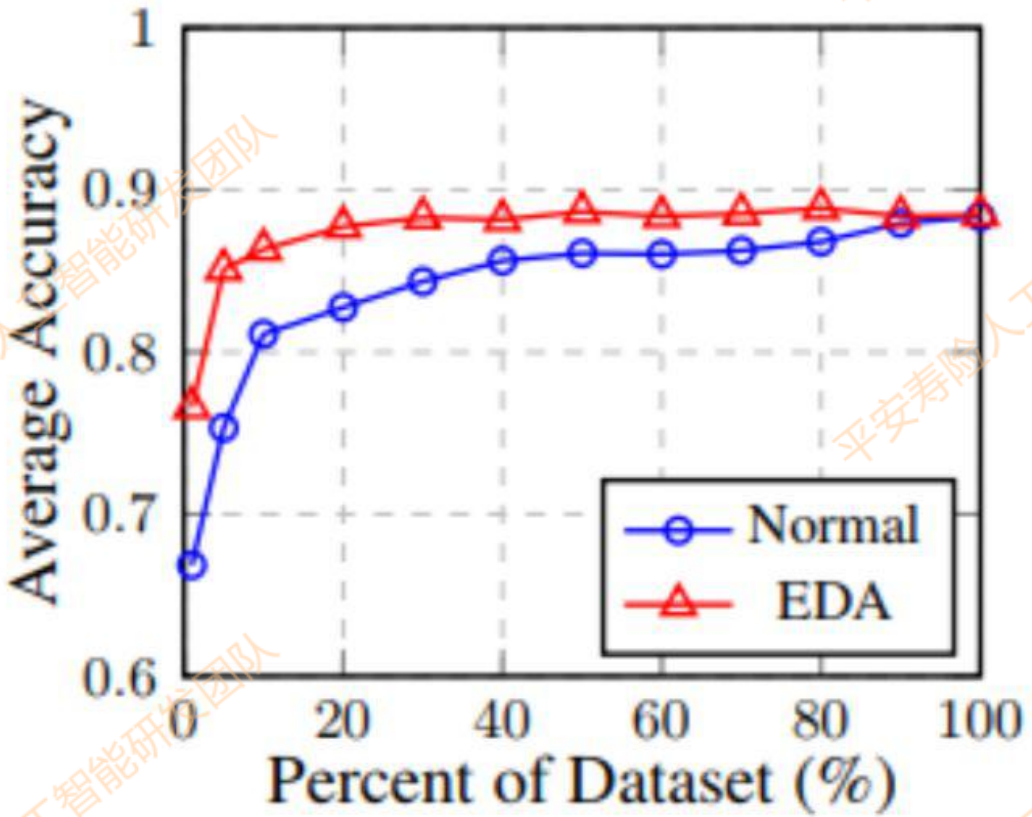
Eg: "如何评价 2017 知乎看山杯机器学习比赛?" —> "2017 机器学习?如何比赛知乎评价看山杯”

(4) 随机删除(RD: Randomly Delete):

句子中的每个词，以概率p随机删除。

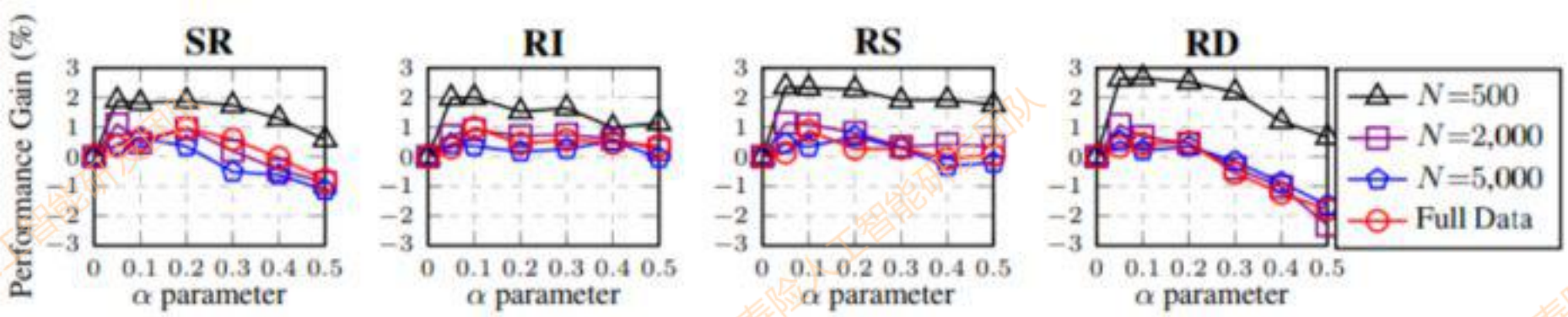
Eg: "如何评价 2017 知乎看山杯机器学习比赛?" —> "如何 2017 看山杯机器学习 “.

2.1.2 EDA效果（总）



这四种方法的效果如何呢？
在英文的数据上效果很可观。
经过上述四种操作，
数据增强后的句子可能不易理解，
但作者们发现模型变得更加鲁棒了，
尤其是在一些小数据集上。

2.1.2 EDA效果（分）



上图是针对不同训练集大小的五个文本分类任务的EDA操作的平均性能增益。

α参数粗略地表示“每次扩充改变的句子中单词的百分比。”纵轴是模型增益。

我们可以看到，当α = 0.1时，模型提升就能达到很好的效果。

训练数据越少，提升效果效果越明显。

2.1.3 EDA存在的问题

共同的问题：

1、对词作随机处理，没有侧重关键词，还包含了停用词，产生的数据可能语义价值较少。
比如：‘我的花很漂亮’ - ‘地我的花很漂亮’

3、句式的泛化能力较弱，交换顺序、改变语义结构并不能产生正确的新句式。

4、随机插入、随机删除、随机交换会产生语病句，不适用于序列模型。

同义词替换SR: 同义词具有非常相似的词向量，而训练模型时这两个句子会被当作几乎相同的句子，而在实际上并没有对数据集进行有效的扩充。

随机插入RI、随机删除RD ：语义逻辑有可能改变，和原标签不一致的错误率较高。

随机交换RS: 实质上并没有改变原句的词素，对新句式、句型、相似词的泛化能力实质上提升很有限。

2.2.1 回译简介

在这个方法中，我们用机器翻译把一段中文翻译成另一种语言，然后再翻译回中文。

Eg: "周杰伦是一位华语乐坛的实力唱将，他的专辑卖遍了全球" —>

"Jay Chou is a strength singer in the Chinese music scene, his albums are sold all over the world. "—>

“周杰伦是中国音乐界的优秀歌手，他的专辑畅销全世界。”

这个方法已经成功的被用在Kaggle恶意评论分类竞赛中。反向翻译是NLP在机器翻译中经常使用的一个数据增强的方法。其本质就是快速产生一些翻译结果达到增加数据的目的。

2.2.2 回译优缺点

优点：

回译的方法往往能够增加文本数据的多样性，相比替换词来说，有时可以改变句法结构等，并保留语义信息。

缺点：

1、回译的方法产生的数据依赖于翻译的质量，大多数出现的翻译结果可能并不那么准确。

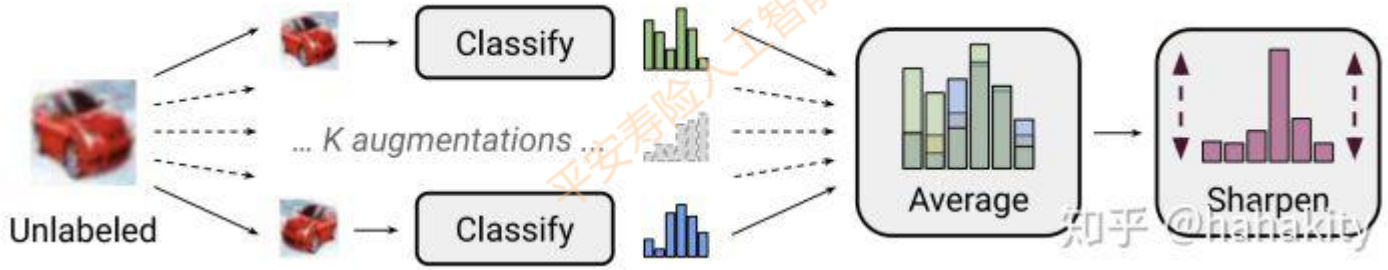
2、中文和外语的同义词数量不同，且模型会倾向寻找正确率更高的词，泛化能力可能受限。（比如：中文音箱、音响、扬声器是同义词，在英文中最常用的都是speaker）

3、如果使用某些翻译软件的接口，也可能遇到账号限制等情况。

三、 深度学习数据增强技术

3.1半监督 Mixmatch

它的工作方式是将有无标签的数据都进行扩增，把无标签的数据输入分类器得到平均分类概率，应用sharpen算法得到猜测标签。通过 MixUp 把无标签数据和有标签数据混合起来，对增广后的数据训练计算损失项。



这种 MixMatch 方法在小数据上做半监督学习的精度，远超其他同类模型。比如，在CIFAR-10 数据集上，只用250个标签，他们就将误差减小了4倍（从38%降到11%）。在STL-10数据集上，将误差降低了两倍。

3.2.1无监督数据增强UDA结构

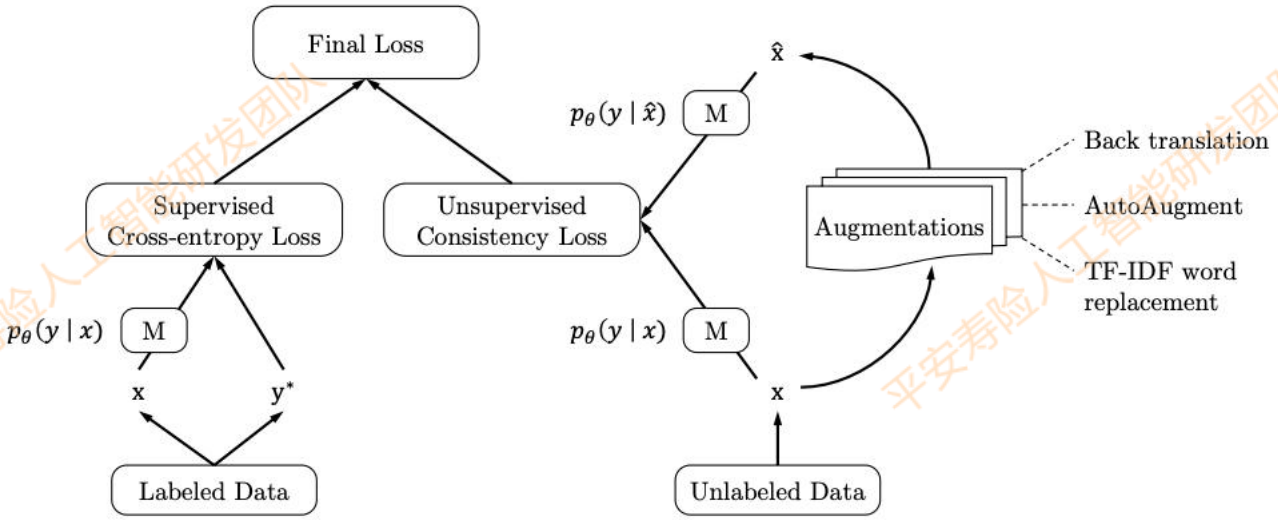


Figure 1: Training objective for UDA, where M is a model that predicts distribution $p_\theta(y \mid x)$ given x, and y^* is the ground-truth label.

UDA 的成功，得益于对特定任务使用**特定目标的数据增强算法**。

在文本的处理方式上论文选用了**回译**和**关键词提取**两种方式。

回译：丰富数据的句式和句型

Tfidf：优化了EDA的随机处理词策略，根据DBPedia先验知识和实际语料的词频确定关键词，再根据确定好的关键词替换同义词，避免无用数据和错误数据的产生。

3.2.2 Training Signal Annealing（TSA）

UDA优秀的另一个重要的突破是采用了**Training Signal Annealing（TSA）训练信号逐步减退**方法在训练时逐步释放训练信号。

因为需要采用大量的未标记数据进行训练，所需的模型会偏大，而大模型又会轻松的在有限的有监督数据上过拟合，这时TSA就要逐步的释放有监督数据的训练信号了。

作者对每个training step 都设了一个阈值ηt，且小于等于1，当一个标签例子的正确类别P的概率高于阈值ηt时，模型从损失函数中删除这个例子，只训练这个minibatch下其他标记的例子。

3.2.2 Training Signal Annealing（TSA）

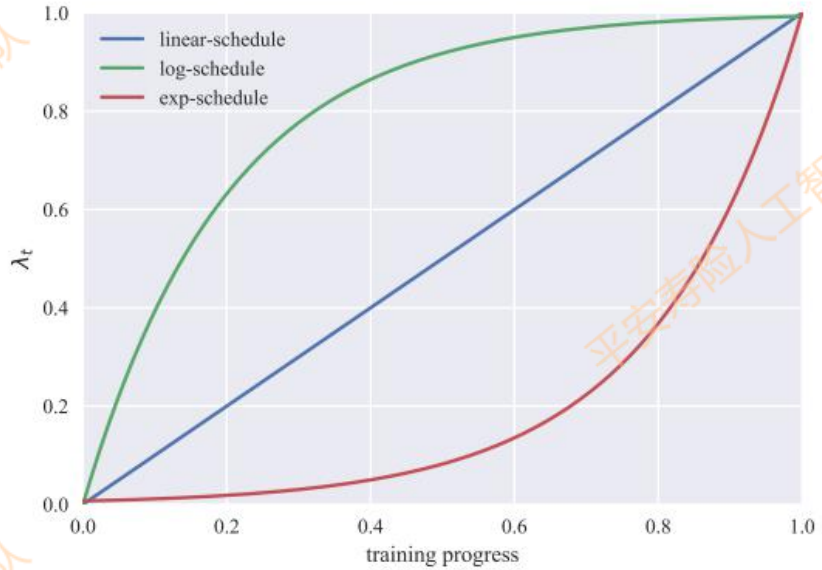
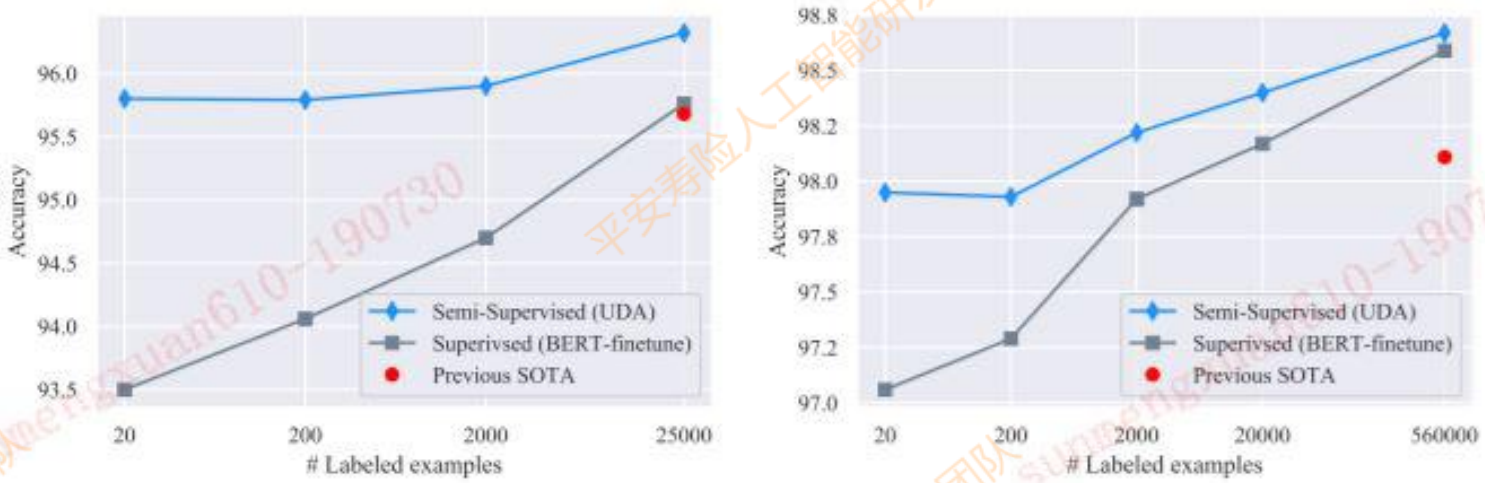


Figure 2: Three different schedules of TSA where λ_t is increased from 0 to 1. We simply set $\eta_t = \frac{1}{k} + \lambda_t * (1 - \frac{1}{k})$ so that η_t goes from $\frac{1}{k}$ to 1.

如上图展示了3种TSA的方式，
这3种方式适用于不同数据。
exp模式更适合于问题相对容易
或标注量较少的情况。因为
监督信号主要在训练结束时释放，
且可以防止模型快速过拟合。
同理，log模式适合大数据量的情况，
训练过程中不太容易过拟合。

3.2.3 UDA效果

作者的实验结果显示，这种无监督方法创造的数据在多个任务上都有很好的表现，甚至在 IMDb 数据集的分类测试中，UDA 只使用 20 个标签就得到了比此前最好的方法在 25,000 个有标签数据上训练更好的结果。



(a) IMDb　　(b) Yelp-2

Figure 4: Accuracy on IMDb and Yelp-2 with different number of labeled examples.

四、数据增强技术实践

4.1 某歌翻译软件回译：

原句：生活里的惬意，无需等到春暖花开

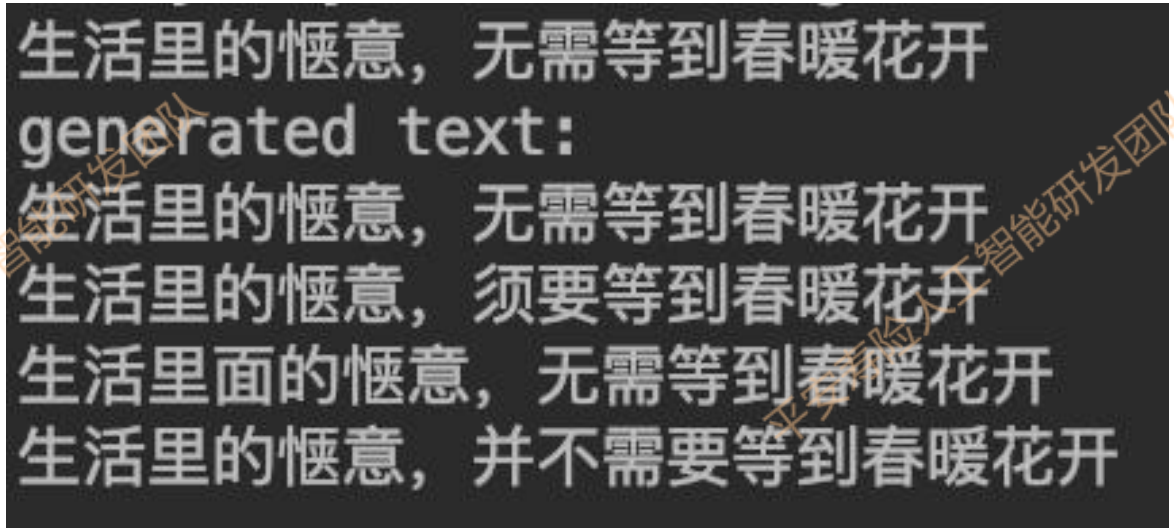
中—>英—>中：生活的舒适，无需等到春天开花

中—>日—>中：生活的舒适，无需等到春天的花朵

中—>德—>中：生活的舒适，无需等到春天开花

中—>法—>中：生活的舒适，无需等待春天的花朵

4.2 传统文本数据扩增产生的数据：



地址：http://github.com/wac81/textda （代码已打包开源为textda：eda+回译）
调用方式：

```
pip install textda
from textda.data_expansion import *
print(data_expansion('生活里的惬意，无需等到春暖花开'))
```

4.3 传统方法对不平衡文本分类的效果提升

此处以情感正中负文本3分类结果为例：

最初训练文本：neg1468，pos 8214， neu 712

测试文本：neg1264，pos 1038，neu 708

分类模型：fastText文本分类器训练模型

由下图的confusion matrix 可知模型整体加权 f1值为 0.749

利用传统的方法将数据扩充至

neg:7458 ， pos:8214 ， neu:3386

当数据趋于平衡，f1值又迎来上升到0.783，

将近4个百分点

Confusion matrix, without normalization

	precision	recall	f1-score	support
neg	0.88261	0.79114	0.83438	1264
neu	0.74384	0.42655	0.54219	708
pos	0.67233	0.95279	0.78836	1038
micro avg	0.76113	0.76113	0.76113	3010
macro avg	0.76626	0.72350	0.72164	3010
weighted avg	0.77746	0.76113	0.74978	3010

Confusion matrix, without normalization

	precision	recall	f1-score	support
neg	0.85425	0.83465	0.84434	1264
neu	0.75264	0.50282	0.60288	708
pos	0.74731	0.93738	0.83162	1038
micro avg	0.79203	0.79203	0.79203	3010
macro avg	0.78474	0.75829	0.75961	3010
weighted avg	0.79347	0.79203	0.78316	3010

数据增强的作用

1、增加数据，提供模型训练

2、防止模型过拟合

3、降低标注成本

4、减少分类时数据不平衡的预测偏差

五、数据增强的拓展

5.1 其他数据增强方法

（1）、音频：

噪声增强

随机相同类型抽取拼接

时移增强

音高变换增强

速度调整

音量调整

混合背景音

增加白噪声

移动音频

拉伸音频信号

（2）、图像：

水平翻转垂直翻转

旋转

缩放 放大缩小

裁剪

平移

高斯噪声

生成对抗网络 GAN

AutoAugment

（3）、其他文本数据增强方法：

语法树结构替换

篇章截取

seq2seq序列生成数据

生成对抗网络 GAN

预训练的语言模型

5.2 防止过拟合其他方法

(1)、**Regularization**：数据量比较小会导致模型过拟合，使得训练误差很小而测试误差特别大. 通过在Loss Function 后面加上正则项可以抑制过拟合的产生。缺点是引入了一个需要手动调整的hyper-parameter。

(2)、**Dropout**：这也是一种正则化手段，不过跟以上不同的是它通过随机将部分神经元的输出置零来实现。

(3)、**Unsupervised Pre-training**：用Auto-Encoder或者RBM的卷积形式一层一层地做无监督预训练，最后加上分类层做有监督的Fine-Tuning。

(4)、**Transfer Learning（迁移学习）**：在某些情况下，训练集的收集可能非常困难或代价高昂。因此，有必要创造出某种高性能学习机（learner），使得它们能够基于从其他领域易于获得的数据上进行训练，并能够在对另一领域的数据进行预测时表现优异。

六、总结和展望

文本数据增强方法

一、传统：

1、EDA：

- 同义词替换（SR：Synonyms Replace）、
- 随机插入(RI：Randomly Insert)、
- 随机交换(RS：Randomly Swap)、
- 随机删除(RD：Randomly Delete)

2、回译

二、半监督深度学习：

1、Mixmatch

2、Uda

文本数据增强的关注点

（1）增加的数据要保证和原数据一致的语义信息。

（2）增加的数据需要多样化。

（3）增加的数据要避免在有标签数据上过拟合。

（4）增加的数据和原数据保持一定的平滑性会更有价值，提高训练效率。

（5）增加数据的方法需要带着目标去选择。