

# 使用卷积神经网络进行单词难度预测

Arpan Basu<sup>\*</sup>, Avishek Garain<sup>†</sup>, Sudip Kumar

Naskar<sup>§</sup> Jadavpur 大学计算机科学与工程系  
印度加尔各答

<sup>\*</sup> arpan0123@gmail.com <sup>†</sup> avishekgarain@gmail.com <sup>§</sup> sudip.naskar@gmail.com

## 摘要

大多数文本简化系统需要一个词的复杂性指标。目前流行的词语难度预测方法是基于人工特征工程。基于深度学习的模型由于其相对较差的性能而在很大程度上没有被探索。在本文中, 我们探讨了在预测单词难度时使用其中的一种方法。我们把这个当做一个二元分类问题。我们训练传统的机器学习模型并评估它们在该任务上的表现。消除对先前获得的单词的频率的依赖以测量难度是我们的主要目的之一。然后, 我们分析了一个基于卷积神经网络的预测模型, 该模型在字符水平上运行, 并评估了它与其他模型相比的效率。

**Index Terms**-word-difficulty, character-CNN, logistic regression, random forest classifier, support vector machine

## I. 简介

在大多数自然语言处理(NLP)任务中, 单词往往是处理的基本单位。研究人员经常使用从词中提取的特征作为计算系统的重要组成部分来解决许多NLP任务。在特别强调文本简化的情况下, 通常的目标是用简单的词来替代文本中存在的难词。这样的系统通常需要对单词的难度进行一些衡量。

然而, 基于模型的方法在很大程度上没有被用来预测单词的难度。部分原因是它们在对某一特定词的有限特征的情况下产生的结果很差。当使用一个模型来预测一组固定的词的连续价值的难度时, 这是可以预期的。这个问题也可以被修改, 但会损失一些信息, 变成一个二元分类问题。我们需要预测该词属于简单和困难两类之一的概率。

在这份报告中, 我们建立并评估了基于上述分类问题的模型。我们还提出了一个基于字符级卷积的单词难度预测模型。

本文其余部分组织如下。第二部分包括对该领域工作的相关研究的简要概述。第三部分描述了数据集和为准备输入数据所做的修改。第四部分包括实验设置和对所使用的方法的一些简要讨论。此后, 第五节显示并阐述了结果。最后, 我们在第六节中总结了本报告。

## II. 相关作品

在英语段落的可读性领域已经有了大量的工作, 从判断军事手册的难度到哈利波特书籍的可读性, 都有自己的应用。在这些衡量标准中, 有一个常用的可读性衡量标准是Flesch-Kincaid评分[6]。它对于段落和基于上下文的文本非常有效, 但是这个分数不能应用于单个单词。因此, 在这种情况下, 必须考虑其他方法。

目前, Kuperman等人的习得年龄评级。[9]提供了一个关于单词难度的良好指示。在这种情况下, 单词的其他各种特征也被使用。例如, 词频和词长已经被成功地用于提供对单词难度的近似估计。我们强调Brysbart和New的相关工作[3], 他们使用从SUBTLEXUS语料库中提取的词频。这里特别指出, 基于词频的模型需要基于一个至少包含2000万个词的语料库。收集这样规模的语料库本身就是一个具有挑战性的问题。因此, 我们需要找到一些替代方案。

Keskisarkka和Robin的自动文本简化[5]利用了同义词替换。这种同义词的重新安置需要识别单词的难度。用同义词替换容易和困难的单词可能会导致效率的降低。我们的方法可能有助于上述的使用情况。

由于在这一领域的工作有限, 我们在收集数据和寻找解决我们在应用这一方法时所面临的问题方面遇到了一些困难。

## III. 数据集

为了训练各种模型, 我们需要一个具有相应难度测量的单词列表。为此, 我们使用了作为英语词典项目[1]的一部分的可用数据。它衡量的是对某一特定单词进行词义判断所需的时间。特别是, 我们使用整个数据集中的 *I Zscore* 和 *Freq* *HAL* 特征。 *I Zscore* 是每个词的平均词汇决策延迟。 *Freq* *HAL* 指的是基于1.31亿个单词的HAL语料库的Hyperspace Analogue to Language频率规范[10]。我们将一个词的难度与 *I Zscore* 直接联系起来。较高的分数代表较高的词性

决策时间，因此是一个难度较高的词，反之亦然。我们还使用 *Freq* - *HAL* 值来构建以特定单词的频率为特征的模型。

我们认为这个问题是一个二元分类的问题，并根据词汇的决策时间将词汇分解为两个类别--0代表 "简单", 1代表 "困难"。我们

选择一个任意的阈值为0，产生两类词。形式上，如果  $I Zscore \leq 0$ ，该类为0，如果  $I Zscore > 0$ ，该类为1。

方式的结果是大体上平等地分配字数——。0班有22621个单词，1班有17828个单词。

#### IV. 实验设置

我们构建了四个模型来测试我们对问题的重新表述。

- A. Logistic Regression (LogReg)
- B. 支持向量机 (SVM) [4]
- C. 随机森林分类器 (RFC) [2]
- D. 字符CNN(CharCNN) [13]

前三个基于机器学习的模型是基于目前使用频率和词长作为特征的方法。它们共享相同的输入特征、输入标签和输出，只在基础算法上有所不同。输入

由一个  $40449 \times 2$  的矩阵组成，其中每一行表示字。每一列条目由两个元素组成--第一个是 *HAL* 频率，第二个是单词的长度。相应的输入标签被提供为

$40449 \times 1$  向量，其中每个元素要么是0，要么是1，取决于对应词的类别。这些模型输出一个  $40449 \times 1$  的向量，根据预测的类别，每个条目都是0或1。我们用准确率来比较这些结果  
公制。

##### A. 逻辑回归

Logistic回归是一个常用的机器学习模型。与线性回归不同，Logistic回归使用Logistic sigmoid函数来返回一个概率值，然后可以用来映射两个（或多个）类别。

##### B. 支持向量机

支持向量机是一种有监督的机器学习模型。它是一个基于生成最佳超平面的判别性分类器，然后用于对输入进行分类。对于二元分类问题，超平面是一条将平面分为两部分的线。

##### C. 随机森林分类器

随机森林分类器是一个决策树的集合体。它根据训练数据的各种随机子样本创建一些决策树。然后，它通过汇总各个决策树的输出来产生输出预测。这减少了单棵树容易受到的变异和噪音。

##### D. CharCNN模型

这个模型是基于一个类似于Zhang等人[13]的字符级卷积模型。我们使用一个大大简化的架构，只应用一个卷积。其工作流程图如图1所示。取出单词，并将其一键编码为21个矢量的序列，每个矢量大小为26。26这个值代表字母表的大小（我们使用所有小写字母），21代表一个词的最大长度。对于一个较短的词，上述序列被填充了零个向量。对于较长的单词，第21个位置之后的字符被忽略。我们注意到，在所使用的数据集中，最长的词是21个字符。最终，这样产生的输入是一个  $40449 \times 21 \times 26$  的矩阵。每个输入词都是一个单热编码的  $21 \times 26$  矩阵。

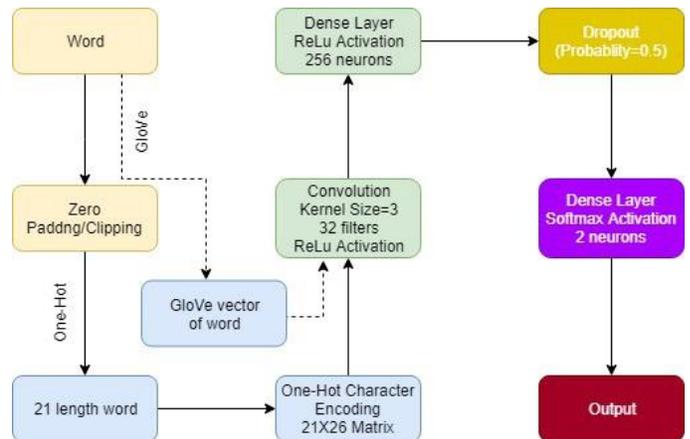


图1.工作流程图

在每个输入词上，我们用32个内核大小为3的滤波器进行一维卷积，并使用ReLU激活。这连接到一个由256个神经元组成的密集层，具有ReLU激活和0.5概率的辍学[12]。密集层与2个具有softmax激活的神经元相连。这2个神经元形成一个词的输出，这个词是  $1 \times 2$  的概率的向量。

第一个值是该词属于0类的概率，第二个值是该词属于1类的概率。考虑到所有的单词，整个输出是一个  $40449 \times 2$  的矩阵。它试图将每个基于字符组的相对位置的单词。这个模型是用Adam[7]的优化算法训练的，批量大小为250，共25个epochs。我们注意到，这个模型没有收到任何与单词的频率或长度相对应的输入。

我们发现，CharCNN模型的表现与基于频率和字长的传统模型相当。即使该模型没有接受频率作为输入特征，而是直接在字符层面上操作，情况也是如此。使用该模型背后的直观推理是

一个词的难度（在合理的程度上）是基于某些连续出现的字符组。卷积网络有效地抓住了这种定位特性，使其能够提供准确的预测。

实验表明，只要不是像 "abcdefghijkl" 那样的任意词，词的确切顺序似乎并不重要。在卷积层之后添加一个时间分布的双向LSTM层会使性能下降1%左右。我们没有报告该模型的结果。此外，在这个模型中加入频率或长度可能会提高其准确性。

### E. 使用预先训练好的嵌入，而不是单热编码

为了进行比较，我们还使用预先训练好的GloVe[11]单词嵌入作为输入特征，而不是单热编码，来评估字符CNN模型。该模型

该模型的基本结构保持不变。每个单一的输入现在被视为一个 $300 \times 1$ 的矩阵，而不是一个 $21 \times 26$ 的矩阵，一维卷积被应用在

它。该模型的这一变体在表中被称为*GloVe*。尽管有更好的准确性和更有希望的结果，但这些嵌入的明显缺点是需要一个大的语料库和不能推断出语料库中没有的词。

### F. 皮尔逊相关

两个变量之间的线性关系的程度是由佩尔森相关性给出。它是一个一般在-1到+1之间的数值。

如果数值为0，则表示与"我"之间没有关联。这两个变量。

大于0的数值表示正相关；也就是说，随着一个变量的数值增加，另一个变量的数值也会增加。

在我们计算这个数值的方法中，我们使用了以下公式。

$$\text{Cov}(X, Y)$$

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

其中

Cov是协方差

$\sigma_X$  是X的标准偏差  $\sigma_Y$

是Y的标准偏差  $X = 40449 \times 1$ 的矢量。

$$x_i = P(w_i \in \text{类1}) - P(w_i \in \text{类0})$$

$$Y = 40449 \times 1 \text{ 矢量, 包含每个 } IZScore$$

## V. 结果

表一  
表中显示了完整数据集的准确率(%)。

日志Reg	证券公司	RFC	剑桥新闻网	AAA
73.98	74.30	86.40	80.10	95.93

表二  
表中显示了10倍交叉验证的准确性(%)。

M=平均值, SD=标准偏差

折叠号	日志Reg	证券公司	RFC	剑桥新闻网	
				一击即中	AAA
1	73.21	74.20	71.95	73.68	75.75
2	73.92	75.48	71.57	72.81	77.55
3	75.38	74.36	72.34	72.26	75.70
4	74.68	74.96	72.78	71.40	75.60
5	73.05	74.68	72.24	72.63	76.64
6	74.71	74.16	71.64	73.25	76.59
7	72.93	73.79	72.90	72.76	76.61
8	73.62	73.87	72.68	73.79	76.69
9	74.11	73.37	72.87	73.24	75.57
10	74.16	73.96	72.30	73.19	75.89
<b>M</b>	73.98	74.28	72.33	72.90	76.26
<b>AAA</b>	0.63	0.47	0.39	0.53	0.65

表三  
表中显示了每个交叉验证折叠的皮尔逊系数

折叠号	日志Reg	证券公司	RFC	剑桥新闻网	
				一击即中	AAA
1	0.55	0.54	0.50	0.61	0.67
2	0.54	0.57	0.53	0.62	0.67
3	0.53	0.54	0.51	0.61	0.65
4	0.54	0.54	0.52	0.61	0.66
5	0.54	0.55	0.50	0.60	0.67
6	0.53	0.55	0.51	0.61	0.67
7	0.56	0.57	0.54	0.64	0.65
8	0.53	0.53	0.51	0.61	0.65
9	0.53	0.56	0.51	0.61	0.68
10	0.54	0.55	0.52	0.59	0.66
<b>平均值</b>	0.54	0.55	0.52	0.61	0.66

表四  
词语及其在0类和1类中的概率  
尊敬的

注：\*不存在于ELP数据集中

词语 (w)	$P(w \in \text{类0})$	$P(w \in \text{类1})$
快速	0.96	0.04
迅速的	0.80	0.20
迅速	0.79	0.21
迅速的	0.88	0.12
迅速	0.30	0.70

## 讨论

我们将这些模型与上述输入相匹配，然后根据它们的分层10倍交叉验证的准确性来评估它们的性能[8]。所有的模型在交叉验证的每个迭代中都被重新训练。得到的结果如表二所示。我们发现，所有的方法都产生了几乎相同的准确性分数。这些分数也一直接近于所考虑的各个模型的平均分数，而且个别分数总是高于百分之七十。

同时，为了衡量与数据集中存在的原始IZscore特征的相似程度，我们找到了数据集与IZscore之间的佩尔森相关系数值。

模型输出和  $I$  Zscore 值。更具体地说。

每个词的两个概率值被转换为一个等于  $P(w \in \text{类}1) - P(w \in \text{类}0)$  的单一值。相关性是用这个值来衡量的，它位于 -1 到 1 之间。这样得到的数值见表三。

我们还报告了让模型在整个数据集上运行时的准确性得分。该分数可以在表 I 中找到。在这种情况下，随机森林分类器和基于 GloVe 的字符 CNN 模型大大超过了数据的适应性。同样的情况也出现在基于单次编码的字符 CNN 模型上，尽管程度较轻但很明显。

CharCNN 模型的主要效用在于它能够从一组合适的替换词中选择一个简单的词。

词。让我们考虑上述问题，我们已经有一组替换词  $R = \{$  "快速"、"迅速"、"快速"、"迅速"、"迅速"、"alacritous"  $\}$ 。我们现在预测的是属于第 0 类或第 1 类的词，在所有的词中替换组。

表四中提到了字符 CNN 模型的输出概率向量。这些概率向量使我们能够选择 "fast" 作为集合中最简单的词。我们也可以很容易地从中找到最难的词 "alacritous"，尽管这个词在最初的训练词列表中并没有。这是基于字符的模型比传统的基于频率和字长的模型的优势之一。这在利用较简单的同义词替换单词的文本简化应用中发现了效用。通常情况下，替换集合中的一个或多个词在训练语料库中是不存在的。

## VI. 总结

从实验结果来看，我们声称基于模型的单词难度估计器在应用于现实世界的案例时表现令人满意，如文本简化。传统的机器学习模型在该任务中确实表现良好；提议的基于卷积的模型也是如此。然而，这些使用词频、词长和词字符的模型并没有利用各种词的发音中存在的语音不匹配。这样的特征可能有助于构建更准确的难度预测模型。

## 参考文献

- [1] D.A. Balota, M. J. Yap, K. A. Hutchison, M. J. Cortese, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The english lexicon project," *Behavior Research Methods*, vol. 39, no.3, pp. 445-459, Aug 2007.[在线]. Available: <https://doi.org/10.3758/BF03193014>
- [2] L. Breiman, "随机森林", *机器学习*, 第45卷, 第1期, 第5-32页, 2001年10月.[在线]. Available: <https://doi.org/10.1023/A:1010933404324>
- [3] M. Brysbaert 和 B. New, "超越 Kucera 和 Francis. *行为研究方法*", 第41卷, 第4期, 第977-990页, 2009年11月, 对目前的词频规范进行批判性评估, 并为美国英语引入新的和改进的词频措施。4, pp. 977-990, Nov 2009.[在线]. Available: <https://doi.org/10.3758/BRM.41.4.977>
- [4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no.3, pp. 273-297, Sep 1995.[在线]. Available: <https://doi.org/10.1007/BF00994018>
- [5] R. Keskisa "rkka", "Automatic text simplification via synonym replacement," 硕士学位论文 Linköping University Linköping University, Department of Computer and Information Science, The Institute of Technology, 2012.
- [6] J.P. Kincaid, R. P. J. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automatic readability index, fog count and flesch reading ease formula) for navy enlisted personnel," 01 1975.
- [7] D.P. Kingma 和 J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.[在线]. Available: <http://arxiv.org/abs/1412.6980>
- [8] R. Kohavi, "交叉验证和自举法对准确性估计和模型选择的研究", 在 *第14届国际人工智能联合会议论文集-第2卷*, ser. IJCAI'95. 美国加州旧金山: Morgan Kaufmann 出版公司, 1995年, 第1137-1143页。[在线]. Available: <http://dl.acm.org/citation.cfm?id=1643031.1643047>
- [9] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 english words," *Behavior Research Methods*, vol. 44, no.4, pp. 978-990, Dec 2012.[在线]. Available: <https://doi.org/10.3758/s13428-012-0210-4>
- [10] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 2, pp. 203-208, Jun 1996.[在线]. Available: <https://doi.org/10.3758/BF03204766>
- [11] J. Pennington, R. Socher, and C. D. Manning, "Glove: 全局向量的单词表示", 在 *自然语言处理的经验方法 (EMNLP)*, 2014年, 第1532-1543页。[在线]. Available: <http://www.aclweb.org/anthology/D14-1162>
- [12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: 防止神经网络过拟合的简单方法", *《机器学习研究》杂志*, 第15卷, 第1929-1958页, 2014年。[在线]. Available: <http://jmlr.org/papers/v15/srivastava14a.html>
- [13] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 649-657.[在线]. Available: <http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf>