面向对外汉语报刊教学的文本难易度分类

邹红建[1]　杨尔弘[1]

([1]北京语言大学应用语言学研究所，北京，100083)

摘　要： 本文介绍了利用文本中通用词的覆盖率和文本长度两个因素，通过计算文本难易度，对大规模文本进行自动初步难易分类。实验发现，综合考虑文本的通用词覆盖率和文本长度比仅根据其中一个因素对文本进行难易度分类效果更佳。针对本次实验所用实验语料，当通用词的覆盖率系数α取0.1，文本长度系数β取0.9时，区分效果最佳。

关键词： 文本难易度　对外汉语报刊教学　通用词覆盖率　文本长度

Text Categorization of Complexity Orienting the Press Teaching of Chinese for Foreigners

Zou Hongjian　Yang Erhong

(Institute of Applied Linguistics, Beijing Language and Culture University, Beijing, 100083)

Abstract: This paper introduces how to calculate the Complexity of texts and how to categorize the texts by calculating the Complexity of texts. The experiment shows that it is better to calculate the Complexity of texts by two factors——the proportions of general vocabularies and the lengths of texts. The texts can be well categorized when the coefficient α equals 0.1 and the coefficient β equals 0.9.

Keywords: the Complexity of texts;　the proportions of general vocabularies ; the Lengths of texts

引　言

对外汉语报刊阅读课使用的教材及其他材料，过去一般由教师积累，文本难易的判断则依赖于教师的语感或者经验。这样难以适应报刊阅读课程材料不断更新和短时间内获取大量材料的要求。因此，利用计算机按难易程度对报刊语料进行分类，从而根据不同的难易度要求迅速提供一定量待选材料，显得特别重要。本文正是面向此任务，尝试用计算机对新闻文本按难易程度进行粗略分类，以供对外汉语报刊阅读课程选择课程材料。

一

1．1　文本难易度

文本的难易度，简单的说就是具体的文本的理解是难还是简单。至于到底有多难，则是一个相对比较的结果，没有绝对的难或简单。随着一个人语言能力与认知能力的提高，同样的文本会变得越来越简单。在不同的人眼中，同样的文本可能难易度不一样。但是面向同样群体，例如汉语非母语的汉语学习者，存在着一个难易度标准来决

作者简介：邹红建，男，1983年出生。北京语言大学应用语言学研究所硕士研究生一年级。E-mail: zouhj97@163.com
杨尔弘，女，1965年出生。博士，北京语言大学应用语言学研究所教授，主要研究方向为语言信息处理。E-mail: yerhong@blcu.edu.cn

定学习到何种程度时应该能理解何种文章，这也是编写教材时选择文本材料的依据。

本文所说的难易度，是指面向对外汉语报刊阅读课程的，文本是否容易理解的程度。

1．2　影响文本难易度的因素分析

从文本来看，难易度由以下一些因素决定：

1，词汇的通用程度（越专业的词覆盖率越大越难）；

2，词汇的使用度（使用度低的词覆盖率越大越难）；

3，文本的长度（一般情况下文本越长越难）；

4，语法结构的复杂度（复杂结构的覆盖率越大越难）；

5，文本的领域（一些领域较之另一些领域更为人熟悉，相应的文本难度会降低）；

5，文本牵涉的语用的文化的因素（这些因素很多时候也增加文本的难度）；

6，作者的行文风格以及文本的题材体裁等。

1．3　本实验采用的区分难易度的文本特征

1，文本长度，可用文本中的字总数来表示，也可以用文本中的词总数来表示，本次实验以单位文本的词总数代表该文本长度；

2，单位文本中通用词汇的覆盖率；

1．4　各领域通用词表

各领域通用词表由北京语言大学应用语言学研究所 DCC 博士研究室史艳岚博士研制，含 4009 项词条。

本试验采用各领域通用词表，主要原因如下：

1，各领域通用词表是在大规模中国主流报纸语料的基础上产生的，对新闻类的文本有较好的代表性。

2，通用词汇不是一成不变，而是动态更新的。各领域通用词表是最新的研制结果，代表了目前的领域通用词汇。

二

2．1　对北京大学出版社出版的《新编汉语报刊阅读教程》的验证：

试验选取北京大学出版社出版的《新编汉语报刊阅读教程》（初级本、中级本、高级本）三本教材中的 70 篇文本，将文本分词、词频统计处理后导入数据库，对文本中的通用词、专用词进行标注后以文本为单位统计。结果发现，初级、中级、高级三级文本除了在文本长度上有显著区别外，在文本的通用词覆盖率上并无明显区分。下表是三个等级 70 篇文本的通用词平均覆盖情况：

表 1

等级	低级	中级	高级
通用词覆盖率	80.0665%	80.5731%	79.6755%

2．2　实验假设：

1，文本的难易度不仅取决于其文本长度，而且取决于其词汇的难易程度。

2，如果将词汇划分为通用词汇与专用词汇，那么文本的词汇的难易程度表现在其通用词会的覆盖率上。也就是说，一篇文章中通用词汇越多，相应的专用词汇越少，那么这篇文章就越简单，反之则越难。

3，综合考虑文本的长度以及文本中通用词的覆盖率，可以比单纯考虑文本长度一个因素更科学地对文本难易度进行区分。

2．3　实验步骤：

2．3．1　随机选取 DCC 动态流通语料库中 2005 年主流报纸文本 2549 篇作为实验语料。其中，《北京青年报》文本 43 篇，《光明日报》515 篇，《人民日报》1313 篇，《中国青年报》678 篇。考虑到对外汉语阅读课程对文本长度的要求，所选取的文本文件大小有所限制，从 2KB 到 6KB。

2．3．2　对实验语料进行分词、词频统计，并将相应数据导入 SQLServer2000 数据库计算。分词采用中科院自动化所的分词软件。词频统计程序用 PERL 语言编写。

2．3．3　根据通用词表标注词条属于通用词还是专用词，并以文本为单位计算通用词频次、专用词频次、文本总词数以及通用词、专用词占整个文本的比例。考虑到一般情况下字母词、阿拉伯数字词比较容易，故未统计

在内。

2．3．4　以单篇文本为单位，计算通用词覆盖率，加权计算，与文本长度加权后的值相加，得到整个文本的难易度。计算公式如下：

通用词覆盖率 ＝ 通用词频次和 / 文本总词数

文本长度值 ＝ （当前文本总词数－最小文本总词数）/（最大文本总词数－最小文本总词数）

文本难易度 ＝通用词覆盖率×α＋文本长度值×β

其中，α、β为相应系数，通过调整其具体值来改变通用词覆盖率与文本长度的相对权值。α＋β＝1。

2．3．5　根据难易度对文本按从大到小顺序进行排序，难易度值越大，说明文本越简单，反之则越难。人工选取一定阈值，相应阈值内的文本属于同一难度等级。

2．3．6　分别设置系数（α，β）为(0, 1)、(0. 1, 0. 9)、(0. 2, 0. 8)、(0. 3, 0. 7)、(0. 4, 0. 6)、(0. 5, 0. 5)、(0. 6, 0. 4)、(0. 7, 0. 3)、(0. 8, 0. 2)、(0. 9, 0. 1)、(1, 0)，在不同区间选取300篇文本，由人工标注难易度等级。将计算机分类的结果与人工标注的结果比对，通过调整系数，得到最佳值。

三

3．1　结果数据

在α，β系数不同取值情况下，文本排序位置会发生很大的变化。为测量不同方式排序下文本位置变化的具体量值，本试验采用一种简单的文本位置偏移值累加的办法。即在α，β系数取（m_1，$1-m_1$）时，按照计算出的难易度给文本排序，并给每个文本一个序号。在α，β系数取（m_2，$1-m_2$）时，根据难易度给文本排序，并给每个文本另一个序号。同一文本的不同序号取差的绝对值并累加，得到两种系数取值条件下文本排序的累加偏移值。该值代表不同系数条件下总的文本位置变化情况。比如，同一篇文本在两种不同排序方式下位置序号分别是3和118，那么，该文本在这两种排序方式下的文本位置偏移值为｜3－118｜＝115。将累加偏移值除以文本数量，则得到文本位置平均偏移值。文本位置平均偏移值说明了平均单个文本发生的不同排序情况下的位置变化。

下面是α，β系数三种不同取值情况下的文本位置平均偏移值。

表2

（α，β）取值1	（α，β）取值2	累加偏移值	平均偏移值（小数后一位）
（1，0）	（0，1）	2224862	872.8
（1，0）	（0.5，0.5）	1754018	688.1
（0，1）	（0.5，0.5）	573074	224.8

从表中可以看出，如果单纯以通用词覆盖率（即α＝1，β＝0）或者单纯以文本长度（即α＝0，β＝1）为标准，则两种排序结果相差很大。如果所有文本根据难易度平均分为三个等级，则平均一篇文本在两种排序方式下位置变化了872.8，几乎相差一个等级（因为整个实验语料文本总数为2549，文本位置调整872.8，近于2549的1/3）。

为了验证不同系数取值对于文本难易度区分的不同效果，本实验从实验语料中，随机抽取300篇的文本进行人工标注。标注等级分初级、中级、高级三个等级（分别用1、2、3表示）。通过观察标注文本在何种系数取值情况下难度相同的文本聚在一起的效果最好，来确定最佳的α，β系数。具体方法是通过给定α，β系数计算文本难易度，并按难易度大小排序。标注语料中1级文本68篇，2级文本119篇，3级文本113篇。按照文本难易度排序，理想的状况应该是从第1篇到第68篇为1级文本范围，从第69篇到第187篇为2级文本范围，从第188篇到第300篇。如果在相应等级范围出现的是其它难度等级的文本，则用该难度等级减去出现范围等级，将三个范围的这些相减的值累加，得到给定系数情况下错误分类数。选择错误分类数最小的分类作为最接近人工标注的分类。

表3

α，β系数	1级范围错误分类得分	2级范围错误分类得分	3级范围错误分类得分	错误分类得分总和

1，0	72	69	93	234
0.9，0.1	37	67	56	160
0.8，0.2	29	54	27	110
0.7，0.3	25	45	20	90
0.6，0.4	22	37	15	74
0.5，0.5	18	32	14	64
0.4，0.6	17	29	12	58
0.3，0.7	14	26	12	52
0.2，0.8	13	26	12	52
0.1，0.9	12	24	12	48
0，1	13	26	13	52

从上表中可以看出，当系数α=0.1，β=0.9时，选择错误分类数最小。因此，本次实验采用该系数。则文本难易度计算公式为

文本难易度＝通用词覆盖率×0.1＋文本长度值×0.9

根据验证的情况，初步把系数α=0.1，β=0.9时难易度分级的几个临界点规定为：

文本难易度>=0.924408883时，难易度等级为1（即初级水平）

文本难易度<0.924408883且文本难易度>=0.729389923 时，难易度等级为2 （即中级水平）

文本难易度<0.729389923时，难易度等级为3（即高级水平）

在系数α=0.1，β=0.9情况下，计算文本的难易度，根据以上的确定的难易度分级的几个临界点，对实验语料进行分类。分类结果如下：

表4

难度等级	文本数量	占实验语料的比例
1级	130	5.10%
2级	1220	47.86%
3级	1199	47.04%

3．2　数据分析

从以上实验及其结果数据可以得出以下结论：

3．2．1　文本长度确实是判定文本难易度的一个显著因素，但是仅仅凭文本长度不能科学区分难易度。通过统计发现，北大版《新编汉语报刊阅读教程》就是只有文本长度的显著差别而无通用词覆盖率的差别，不同等级的文本在词汇的难度上没有显示出差别。

3．2．2　仅仅凭通用词覆盖率也不能有效区分文本。从上表中可以看出，单纯凭通用词覆盖率划分难易度等级，其错误得分甚至远远高于仅凭文本长度所作的难易度划分。

3．2．3　只有综合考虑文本长度和通用词覆盖率两个方面，才能提高文本难易度判定的正确率。

3．2．4　通用词覆盖率给的系数越大，则文本难易度分类的错误越多。这是一个值得思考的问题。在原先的设想中，通用词覆盖率应该对文本难易度有更大的影响。实验中发现这种设想并不正确。原因之一在于，一些领域的文本，例如体育，虽然通用词覆盖率并不高（一般文本长度不大），但是由于其内容为大多数外国留学生所熟悉，因此，一般都认为是1级难度（即初级水平）。因此，不同的领域的文本，在通用词覆盖率相同，文本长度近似的情况下，领域越为人熟悉则难度越低。原因之二在于，本实验是用词总数代表文本长度，在将词总数归一化过程中，其大小除了取决于具体文本的词总数，还取决于实验语料中最大文本长度和最小文本长度。因此，该实验难易度计算公式中α、β的最佳取值不是普遍适用的。

3．3 验证

为验证上述文本难易度分类计算方法的精确性，本次实验从实验语料中随机抽取150篇文本（前面已经标注的300篇文本不在抽取范围之列），人工标注难易度等级并按难易度值将文本从大到小排序。根据前面所得的难易度等级分界点，将150篇测试文本分为1、2、3三个等级。然后分别计算其错误分类数，计算方式同上。结

果如下：

表 5

等级	文本数	错误分类得分数
1 级	30	9
2 级	62	26
3 级	58	2
总和	150	37

人工标注各等级文本数量与计算机错误分类个数如下表：

表 6

等级	人工标注的文本数	分类不一致文本数	正确率
1 级	39	11	71.8%
2 级	47	17	63.8%
3 级	64	6	90.6%
总和	150	34	77.3%

四

本次实验对文本难易度自动分类作了一些探讨，还有很多值得进一步思考、深化的地方。下面是进一步的工作设想：

4．1　文本难易度不仅取决于所用的词语，还取决于语法结构，以及语用因素和文化因素。在后面的实验中，可以考虑增加语法结构的因素，以《汉语水平语法等级大纲》所规定的语法结构为判断标准，综合考虑文本长度、词汇和语法方面的因素。

4．2　引进文本领域度的概念及计算。如前所述，在人工标注文本难易等级的过程中，文本所属领域也是一个重要的考虑因素。因此，对待不同领域的文本，应该考虑通过文本领域度的系数进行调整，使根据难易度计算的分类结果更接近于人的标注结果。

4．3　由于时间限制，本次实验所用的实验语料规模比较小，标注难易度的文本数量也比较小。在进一步的工作中，将扩大实验语料的选择范围以及数量，标注难易度的文本数量也将进一步扩大，以求获得一个更为科学精确的结果。

4．4 如前所述，本实验是用词总数代表文本长度，在将词总数归一化过程中，其大小除了取决于具体文本的词总数，还取决于实验语料中最大文本长度和最小文本长度。因此，该实验难易度计算公式中 α 、β 的最佳取值不是普遍适用的。如何得到普遍适用的 α 、β 的最佳取值，也是下一步要进行的工作。

4．5 本次实验还考虑了是否采用词汇密度作为判断难易度的一个因素。理论上，词汇密度是衡量单位文本信息量的一个尺度，也是文本难易的一个因素。传统的计算方法是，用文本词项数量除以文本的词总数，再乘以 100%。实验发现，单纯根据词汇密度排序，采用前文所说的计算错误分类的得分为标准，对于实验语料来说，效果甚至没有单纯根据通用词覆盖率分类好。这也是一个值得思考的问题。本次实验由于时间限制，未深入研究这个问题，在进一步的实验中，将对该现象作深入研究。

致谢

本文的实验及写作得到北京语言大学应用语言学研究所史艳岚博士的热心支持，300 篇试验文本及 150 篇验证文本的难易度等级由史艳岚博士人工标注。在此表示谢意！

参考文献

[1]　陈学斌．选用阅读材料的思路和方法[J]．陕西师范大学学报（哲学社会科学版），1997 年 10 月，第 26 卷

[2]　Christopher D．Manning ，Hinrich Schütze．统计自然语言处理基础[M]．2005 年 1 月，北京：电子工业出版社．P355~p374

宴生宏，黄莉．英文易读度测量程序开发探索[J]．重庆大学学报（社科版），2005 年第 11 卷第 2 期