SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON
MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 3, ISSUE 3
ISSN 2054-7390

Difficulty-Level Classification for English Writings

[1]Hiromi Ban, [2]Rei Oguri and [3]Haruhiko Kimura
[1]Graduate School of Engineering, Nagaoka University of Technology, Niigata, Japan;
[2,3]Graduate School of Natural Science and Technology, Kanazawa University, Ishikawa, Japan;
je9xvp@yahoo.co.jp; oguri@blitz.ec.t.kanazawa-u.ac.jp; kimura@blitz.ec.t.kanazawa-u.ac.jp

ABSTRACT

The popularity of e-books has grown recently. As the number of e-books continues to increase, the task of categorizing all books manually requires a significant amount of time. If English sentences can be categorized according to their level of difficulty, it becomes possible to recommend a foreign-language book compatible with the reader's level of competency in English. This study extracted eleven types of attribute from English text data, with the aim of classifying English text according to level of difficulty by learning and categorization. Using the method of "leave-one-out cross-validation," text was subjected to machine learning and categorization. In order to improve accuracy, furthermore, an experiment was carried out in which the size of text data was varied, and the attribute selection method was implemented. As a result, accuracy was improved to 77.04%, and F-measure to 63.96%.

Keywords: Accuracy; Difficulty-level; F-measure; Machine learning.

1 Introduction

The popularity of e-books has grown recently, with the number of books and magazines distributed within Japan in 2014 growing by 18.3% compared with the previous year to 720,000, as shown in Figure 1. Furthermore, it is predicted that in 2016, this number will reach 1.2 million [1].

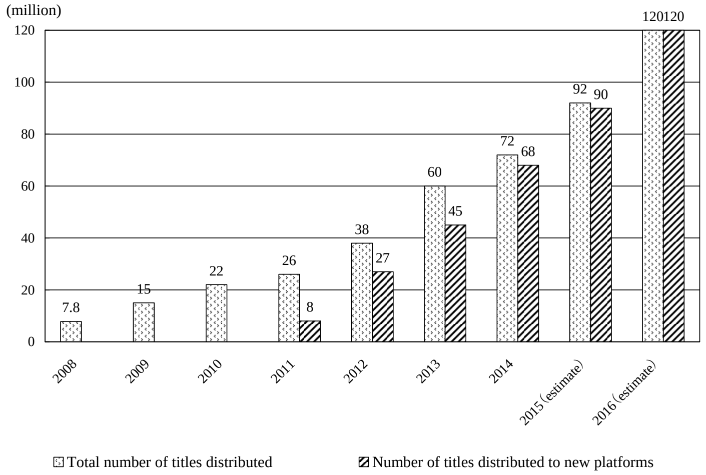


Figure 1. Number of titles of digital books and magazines distributed in Japan.

DOI: 10.14738/tmlai.33.1245
Publication Date: 16th June, 2015
URL: http://dx.doi.org/10.14738/tmlai.33.1245

The number of books listed in the Kindle store as of 28th January 2015 is shown in Table 1, broken down by genre [2]. Compared with the 23 genres of domestically published e-book, all non-Japanese books (of which 3 million are available) are categorized in a single genre.

Table 1. Number of books per genre at Kindle store on Jan. 28, 2015.

Genre	Number	Genre	Number	Genre	Number
Literature & commentary	60,912	Medicine & pharmacology	2,094	Language study, dictionary, cyclopedia & yearbook	1,849
Humanities & thought	17,663	Computer & IT	3,959	Education, study-aid book & examination	3,239
Society & politics	9,118	Art, construction & design	3,209	Picture book & children's book	3,228
Nonfiction	2,611	Hobby & practical use	9,441	Comic	99,187
History & geography	7,854	Sports & outdoor amusement	2,237	Light novel & BL	24,629
Business & economic	11,329	Qualification & authorization	640	Entertainment	2,317
Investment, finance & company management	3,593	Living, health & child-rearing	9,654	Adult	16,912
Science & technology	8,757	Travel guide & map	2,890	Kindle foreign book	3,071,739

As the number of e-books continues to increase, the task of categorizing all books manually requires a significant amount of time; this time requirement becomes even greater if the genre of the book is not clear from its title or the name of its author. In addition to categorization by genre, books may also be categorized according to their level of difficulty. Readers who are studying English may wish to read a simple foreign-language book, while those wishing to extend their language abilities may wish to read a slightly more difficult book. In such cases, analysis is simple, because e-books are a form of electronic data. If English sentences can be categorized according to their level of difficulty, it becomes possible to recommend a foreign-language book compatible with the reader's level of competency in English. For this reason, this research aims to identify the difficulty level of English text.

2 Related Research

In a prior report, the authors implemented quantitative linguistic analysis on English language textbooks used in Finland, which is considered to have the highest level of reading comprehension, mathematical and scientific literacy according to the Organization for Economic Cooperation and Development (OECD)'s Program for International Student Assessment (PISA), and English language textbooks used in Japan, and compared their difficulty level based on the words occurring therein [3]. We also extracted attributes such as the average word length and number of words per sentence.

In this study, the text data and attributes from our previous report were used with the aim of identifying level of difficulty within English sentences.

3 Method

3.1 Data Used

In this paper, the text data used was the same as that used in other related studies, in other words, the textbook used in in third and fourth grade elementary school English lessons in Finland [3][4].

- *Wow! 3* (2002, WSOY)
- *Wow! 4* (2003, WSOY)
- *Wow! 5* (2005, WSOY)
- *Wow! 6* (2006, WSOY)

3.2 Proposed Method

Attributes are extracted from the text data to create data sets. The data sets thus created are subjected to machine learning and categorized.

3.2.1 Attribute Extraction/Data Set Creation

The attributes used for data set creation in this study are the eleven types shown in Table 2.

Table 2. Attributes to be educed.

Total number of characters	Mean word length
Total number of character-type	Words/sentence
Total number of words	Sentences/paragraph
Total number of word-type	Words/word-type
Total number of sentences	Commas/sentence
Total number of paragraphs	

There are a total of 12 objective variables, consisting of grades three through six divided into the three categories of preliminary, intermediate and final phases. This takes into account the fact that even within the same school year, the sentences in the first pages of the textbook have a different difficulty level to those in the final pages.

The eleven attributes were extracted from each text file, and defined as one instance. Table 3 depicts the data sets where as an example, the quantity of text per instance was defined as one page of the textbook.

Table 3. Data set in the case of 1 page per instance.

Total num. of characters	Total num. of character-type	Total num. of words	· · ·	Sentences/ paragraph	Words/ word-type	Commas/ sentence	Class
207	36	40	· · ·	1.25	1.429	0.10	a
252	40	44	· · ·	1.00	1.257	1.17	a
213	37	38	· · ·	1.60	1.226	0.75	a
252	37	52	· · ·	2.00	1.529	0.60	a
261	36	60	· · ·	2.60	1.429	0.08	a
·	·	·	· · ·	·	·	·	·
·	·	·	· · ·	·	·	·	·
·	·	·	· · ·	·	·	·	·
1040	50	181	· · ·	2.57	1.361	0.44	l
1315	58	241	· · ·	2.33	1.461	0.54	l
1526	52	288	· · ·	2.25	1.834	0.44	l
2099	58	396	· · ·	2.04	2.052	0.38	l
2132	54	416	· · ·	1.96	2.286	0.44	l

3.2.2 Machine Learning

The data sets were subjected to machine learning and categorization. Leave-one-out cross-validation was used in learning. Leave-one-out cross-validation is a learning method involving taking one piece of data from the whole as test data, and defining the rest as learning data, and repeatedly validating so that each piece of data becomes the test data once.

The classifier used was a Random Committee.

The classifier used the open source data mining tool Weka in learning and identification [5].

4 Experimentation

In this study, two experiments were carried out using the following evaluation methods during machine learning.

4.1 Evaluation Methods

The evaluation procedure used in this study is as shown in Figure 2.

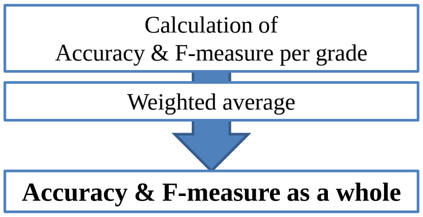

Calculation of
Accuracy & F-measure per grade
Weighted average
Accuracy & F-measure as a whole

Figure 2. Evaluation procedure.

For example, among data predicted by the classifier to be in the fourth-grade textbook, data that actually was in the fourth-grade textbook was defined as a TruePositive, while that not in the fourth-grade textbook was a FalsePositive. Among data predicted by the classifier to not be in the fourth-grade textbook, data that was in fact in the fourth-grade textbook was defined as a FalseNegative, while that not actually in the fourth-grade textbook was defined as a TrueNegative. The threat scores of these categories are compiled in a categorization table such as that in Table 4.

Table 4. Contingency table.

	Correct answer +	Correct answer -
Estimate +	TruePositive	FalsePositive
Estimate -	FalseNegative	TrueNegative

All data was categorized, as in Figure 3, using the 12 objective variables.

			Correct answer											
			3rd grade			4th grade			5th grade			6th grade		
			a	b	c	d	e	f	g	h	i	j	k	l
Estimate	3rd grade	a	8	2	3	2	1	2	0	1	0	0	0	0
		b	3	2	2	3	4	2	0	0	0	0	2	1
		c	1	3	1	4	2	3	2	0	2	1	0	0
	4th grade	d	2	5	3	4	7	3	0	1	0	1	0	0
		e	1	2	3	3	2	1	1	2	0	0	0	1
		f	0	1	1	3	3	6	2	0	1	4	1	0
	5th grade	g	0	1	2	0	0	2	4	1	1	2	2	4
		h	0	0	0	1	1	0	3	2	3	5	7	3
		i	0	1	1	1	0	0	1	5	6	3	2	2
	6th grade	j	2	0	0	1	0	1	4	4	1	4	3	6
		k	0	0	0	0	0	0	3	5	5	4	4	8
		l	0	0	0	0	1	1	3	2	3	6	9	4

	Correct answer +	Correct answer -
Estimate +	TruePositive	FalsePositive
Estimate -	FalseNegative	TrueNegative

Figure 3. Evaluation Method

In addition to the categorization of each academic year into preliminary, intermediate and final phases, the final phase of the previous academic year and preliminary phase of the year above were also counted as correct, giving a total of five correct categorizations for data. In other words, as shown in the example of Figure 3, in the case of the fourth grade textbook, data categorized into either the preliminary, intermediate or final phase of the fourth grade, the final phase of the third grade or the preliminary phase of the fifth grade was considered a correct answer.

The categorization results obtained using the evaluation method shown in Figure 3 were summarized by academic year, as shown in Table 5.

Table 5. Threat score for each grade.

3rd grade	Correct answer +	Correct answer -
Estimate +	35	48
Estimate -	15	173
4th grade	Correct answer +	Correct answer -
Estimate +	43	50
Estimate -	21	148
5th grade	Correct answer +	Correct answer -
Estimate +	38	76
Estimate -	30	127
6th grade	Correct answer +	Correct answer -
Estimate +	55	51
Estimate -	34	131

Next, the rate of accuracy, that is, Accuracy and F-measure were calculated for each academic year.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

Finally, the weighted average was obtained from the calculated accuracy and number of data sets, to calculate overall accuracy and F-measure. This was defined as the evaluation value in this case.

4.2 Experiment 1

4.2.1 Details of Experiment

An experiment was carried out to establish the relationship between changes in the volume of text data used to extract attributes, accuracy and F-measure.

Three types of data set – taking one page, two pages and three pages of text as a single instance of text – were subjected to machine learning and categorization under the conditions shown in Table 6.

Table 6. Experiment environment

Number of characteristics	11
Classifier	Randomcommitte
Technique	leave-one-out cross-validation

The method used to create data sets with two pages of text per instance is as shown in Figure 4.

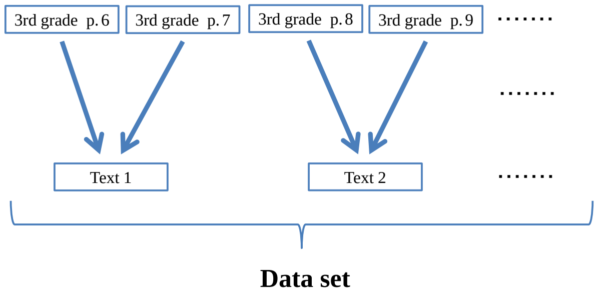


Figure 4. Method of making a data set in the case of 2 pages per instance.

Similarly, with three pages of text per instance, the text data was created in order, so as not to overlap, three pages at a time. The number of instances was 271, 136 and 92, respectively, depending on whether the quantity of text was one, two or three pages.

4.2.2 Results

Results of Experiment 1 are shown in Table 7.

Table 7. Accuracy and F-measure in Experiment 1.

	Accuracy	F-measure
1 page	68.62%	50.95%
2 pages	70.36%	53.48%
3 pages	74.24%	58.87%

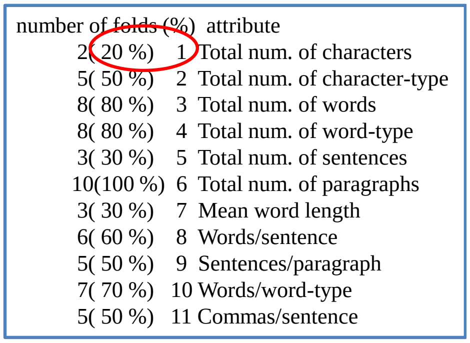
From Table 7 we see that the greater the number of pages, the higher the accuracy and F-measure achieved. Given this, it is considered that using larger quantities of text data for extracting attributes is effective in categorization.

Hereafter, three pages of the textbook will be used per instance when creating data sets for this study.

4.3 Experiment 2

4.3.1 Details of Experiment

The attribute selection method was implemented using the attribute selection function of Weka. The attribute selection method involves searching for items with a low contribution in regard to the objective variable, or attributes that are difficult to predict. These are output as in Figure 5, using attribute selection. The smaller the numerical value, the lower the contribution. A threshold is defined, and attributes below the threshold are deleted, after which attributes are selected once again. Each time attribute selection is implemented, accuracy and F-measure are recorded. This is repeated until all attributes are above the threshold value.

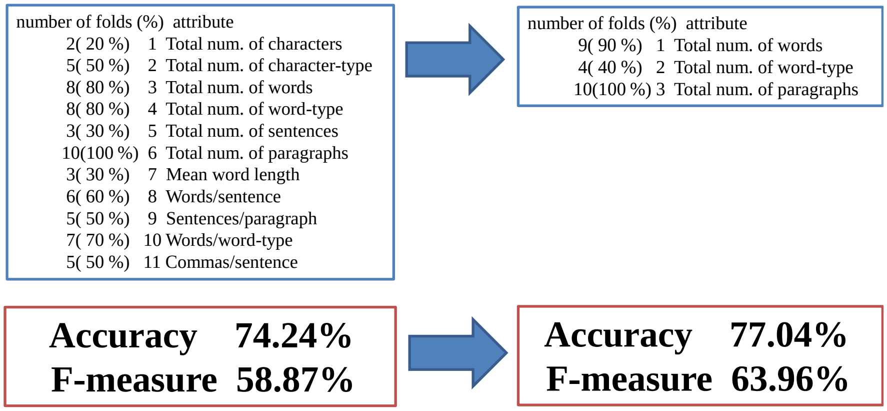


```
number of folds (%)  attribute
         2( 20 %)    1  Total num. of characters
         5( 50 %)    2  Total num. of character-type
         8( 80 %)    3  Total num. of words
         8( 80 %)    4  Total num. of word-type
         3( 30 %)    5  Total num. of sentences
        10(100 %)    6  Total num. of paragraphs
         3( 30 %)    7  Mean word length
         6( 60 %)    8  Words/sentence
         5( 50 %)    9  Sentences/paragraph
         7( 70 %)   10  Words/word-type
         5( 50 %)   11  Commas/sentence
```

Figure 5. Output of feature selection.

4.3.2 Results

After three repeats at threshold value 40%, accuracy and F-measure both demonstrated maximum values. These results are shown in Figure 6.



```
number of folds (%)  attribute
         2( 20 %)    1  Total num. of characters
         5( 50 %)    2  Total num. of character-type
         8( 80 %)    3  Total num. of words
         8( 80 %)    4  Total num. of word-type
         3( 30 %)    5  Total num. of sentences
        10(100 %)    6  Total num. of paragraphs
         3( 30 %)    7  Mean word length
         6( 60 %)    8  Words/sentence
         5( 50 %)    9  Sentences/paragraph
         7( 70 %)   10  Words/word-type
         5( 50 %)   11  Commas/sentence
```

```
number of folds (%)  attribute
         9( 90 %)    1  Total num. of words
         4( 40 %)    2  Total num. of word-type
        10(100 %)    3  Total num. of paragraphs
```

Figure 6. Result of Experiment 2.

As a result, the attribute selection method was implemented, and when the number of attributes was reduced to the following three: “total number of words,” “total number of word types” and “total number of paragraphs,” accuracy increased to 77.04% and the F-measure to 63.9%.

5 Considerations

Accuracy and F-measure were both highest when three pages of text were used per instance. From this, it is believed that the attributes extracted from three pages of text are effective in categorization.

Next, the use of the attribute selection method allowed a reduction in the number of attributes from 11 to 3, and increased accuracy to 77.04% and the F-measure to 63.9%. The remaining three attributes, in other words “total number of words,” “total number of word types” and “total number of paragraphs,” are believed to be those that have the most impact on the difficulty level of English text.

Using these two experiments and reducing the number of attributes improved accuracy, but as shown in Table 8, some data was categorized in significantly erroneous categories.

Table 8. Estimate and correct answer in Experiment 2.

			Correct answer											
			3rd grade			4th grade			5th grade			6th grade		
			a	b	c	d	e	f	g	h	i	j	k	l
Estimate	3rd grade	a	**2**	**0**	**0**	**1**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
		b	**2**	**3**	**0**	**2**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
		c	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
	4th grade	d	**1**	**3**	**2**	**4**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
		e	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**
		f	**0**	**0**	**1**	**0**	**1**	**2**	**2**	**0**	**1**	**0**	**0**	**1**
	5th grade	g	**0**	**0**	**0**	**0**	**1**	**0**	**5**	**1**	**1**	**0**	**0**	**0**
		h	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**3**	**1**	**2**	**1**	**2**
		i	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**2**	**0**	**1**	**1**	**2**
	6th grade	j	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**2**	**0**	**2**	**2**
		k	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**4**	**1**
		l	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**5**	**2**	**2**

When the pages that were significantly mis-categorized were examined, it was found that they all contained columns. In other words, it is believed that the mistaken identification was caused by the impact of the columns between sentences. As a result, it is considered that removing columns from the scope of investigation is likely to improve accuracy.

6 Conclusions

This study extracted eleven types of attribute from English text data, with the aim of classifying English text according to level of difficulty by learning and categorization. Using the method of “leave-one-out cross-validation,” text was subjected to machine learning and categorization. In order to improve accuracy, furthermore, an experiment was carried out in which the size of text data was varied, and the attribute selection method was implemented. As a result, accuracy was improved to 77.04%, and F-measure to 63.96%. At the same time, we noted erroneous identification resulting from the impact of columns between sentences.

In the future, when identifying the difficulty level in English text, we intend to consider new attributes that allow more accurate categorization, and more effective combinations of attribute quantity.

REFERENCES

[1]. ITmedia eBook USER | What is the total number of titles of e-books and e-magazines distributed within Japan? http://ebook.itmedia.co.jp/ebook/articles/1412/19/news033.html

[2]. Kindle Store, http://www.amazon.co.jp/Kindle-%E3%82%AD%E3%83%B3%E3%83%89%E3%83%AB-%E9%9B%BB%E5%AD%90%E6%9B%B8%E7%B1%8D/b?node=2250738051

[3]. Hiromi Ban and Takashi Oyabu, Text Mining of English Textbooks in Finland, "Proceedings of the Asia Pacific Industrial Engineering & Management Systems Conference 2012", V. Kachitvichyanukul, H.T. Luong and R. Pitakaso eds., pp.1674-1679.

[4]. Wow! 3 (2002, WSOY) Wow! 4 (2003, WSOY) Wow! 5 (2005, WSOY) Wow! 6 (2006, WSOY), http://www.kknews.co.jp/developer/finland/

[5]. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/