

2017 9th International Conference on Knowledge and Systems Engineering (KSE)

Sponsored by



TOPICA AILab



IEEE Catalog Number CFP1703I-ART
ISBN 978-1-5386-3576-6



2017 9th International Conference on Knowledge and Systems Engineering (KSE)



IEEE CFP1703I-ART
ISBN 978-1-5386-3576-6



2017 9th International Conference on Knowledge and Systems Engineering (KSE)

Hue, Vietnam, October 19-21, 2017



Editors

Thanh Thuy Nguyen
Anh Phuong Le
Satoshi Tojo
Le Minh Nguyen
Xuan Hieu Phan

Celeberating 60th Anniversary of Hue University of Education

2017 9th International Conference on Knowledge and Systems Engineering (KSE)

KSE 2017

Table of Contents

Preface	vii
Organizing Committee	viii
Program Committee	ix
Keynote Addresses	x
Special Sessions	xiii
Paper Awards	xvi
A new fall detection system on Android smartphone: application to a SDN-based IoT system <i>Hai Anh Tran, Thu Ngo-Quynh and Van Tong</i>	1
An ensemble learning-based method for prediction of novel disease-microRNA associations <i>Duc-Hau Le, Van-Huy Pham and Thuy Thi Nguyen</i>	7
Some observations on representation of dependency degree k <i>Hoang Son Nguyen, Hung Son Nguyen, Long Giang Nguyen and Ngoc Thuy Nguyen</i>	13
Distributed Algorithm for Sequential Pattern Mining on a Large Sequence Dataset <i>Tho Hoang, Bac Le and Minh-Thai Tran</i>	18
Multi-channel LSTM-CNN model for Vietnamese sentiment analysis <i>Quan-Hoang Vo, Huy-Tien Nguyen, Bac Le and Minh-Le Nguyen</i>	24
A Knowledge Representation for Vietnamese Legal Document System <i>Ha-Thanh Nguyen, Viet-Ha Nguyen and Viet-Anh Vu</i>	30
Examining the Text-length Factor in Evaluating the Readability of Literary Texts in Textbooks <i>An-Vinh Luong, Diep Nguyen, and Dien Dinh</i>	36
Prediction-based optimization for online People and Parcels share a ride taxis <i>Son Nguyen Van, Dung Pham Quang, Behrouz Babaki, Hoai Nguyen Xuan and Anton Dries</i>	42
Predicting Students' performance based on Learning Style by using Artificial Neural Networks <i>Binh Hoang Tieu and Duy Bui The</i>	48
DF-SWin: Sliding Windows for Multi-Sensor Data Fusion in Wireless Sensor Networks <i>Huy Duong-Viet and Viet Nguyen-Dinh</i>	54
On the Usage of Character Distribution for the Detection of Web Attacks <i>Thanh Le Dinh and Tien Phan Xuan</i>	60
Construction of a Word Similarity Dataset and Evaluation of Word Similarity Techniques for Vietnamese <i>Tan Bui Van, Thai Nguyen Phuong and Lam Pham Van</i>	65
Sequential Ensemble Method for Unsupervised Anomaly Detection <i>Huy Van Nguyen, Trung Thanh Nguyen and Quang Uy Nguyen</i>	71
Generative Software Module Development: A Domain-Driven Design Perspective <i>Duc Minh Le, Duc-Hanh Dang and Viet-Ha Nguyen</i>	77

A New Approach for Traffic-Sign Recognition using Sparse Representation over Dictionary of Local Descriptors <i>Do Thanh Ha, Nguyen Tien Dat and Le Ngoc Tuan</i>	83
An Empirical Study of Discriminative Sequence Labeling Models for Vietnamese Text Processing <i>Phuong Le-Hong, Minh Quang-Nhat Pham, Thai-Hoang Pham, Tuan-Anh Tran and Dang-Minh Nguyen</i>	88
Fuzzy Relational Compositions Based On Grouping Features <i>Nhung Cao, Martin Stepnicka, Michal Burda and Ales Dolny</i>	94
A Tool Support to Checking Consistency in Model Refactoring <i>Thi Huong Dao, Thanh Binh Trinh and Ninh Thuan Truong</i>	100
Stratifying Cancer Patients based on Multiple Kernel Learning and Dimensionality Reduction <i>Thanh Trung Giang, Thanh Phuong Nguyen and Dang Hung Tran</i>	106
Toward integrating social networks into Intelligent Tutoring Systems <i>Huynh-Ly Thanh-Nhan, Le Huy-Thap and Thai-Nghe Nguyen</i>	112
Single View Image Based – 3D Human Pose Reconstruction <i>Hoang Trung Kien, Nguyen Kim Hung, Ma Thi Chau, Ngo Thi Duyen and Nguyen Xuan Thanh</i>	118
Vietnamese Food Recognition Using Convolutional Neural Networks <i>Van Phat Thai, Tien Dang Xuan, Quang Pham Hong, Nguyen Pham Kieu Thao and Binh Nguyen</i>	124
Facial Expression Recognition Using Deep Convolutional Neural Networks <i>Sang Dinh Viet, Thuan Do Phan and Dat Nguyen Van</i>	130
Facial Smile Detection Using Convolutional Neural Networks <i>Sang Dinh Viet, Thuan Do Phan and Cuong Le Tran Bao</i>	136
Genomedics: Whole exome analysis system for clinical studies <i>Le Sy Vinh, Bui Ngoc Thang, Nguyen Duc Canh, Tran Cong Hoang, Duong Quoc Chinh, Do Thi Thu Hang, Le Ba Hong Minh and Pham Thi Dieu Linh</i>	142
Inconsistency Measures for Probabilistic Knowledge Bases <i>Van Tham Nguyen and Trong Hieu Tran</i>	148
Question Analysis for Vietnamese Legal Question Answering <i>Ngo Xuan Bach, Le Thi Ngoc Cham, Tran Ha Ngoc Thien and Tu Minh Phuong</i>	154
Minimizing the Spread of Misinformation on Online Social Networks with Time and Budget Constraint <i>Manh Vu Minh and Huan Hoang Xuan</i>	160
Density-based clustering with side information and active learning <i>Viet-Vu Vu and Hong-Quan Do</i>	166
Enhanced Semantic Refinement Gate for RNN-based Neural Language Generator <i>Tran Van Khanh and Nguyen Le Minh</i>	172
Multi-Column CNNs for skeleton based human gesture recognition <i>Hai Nguyen, Nam Ly, Huong Truong and Dung Nguyen</i>	179

A Search Method for Ordinances and Rules in Japanese Local Governments Based on Distributed Representation <i>Kazuya Fujioka, Makoto Nakamura, Yasuhiro Ogawa, Tomohiro Ohno and Katsuhiko Toyama</i>	185
Plant Identification using combinations of multi-organ images <i>Thanh Binh Do, Huy Hoang Nguyen, Thi Thanh Nhan Nguyen, Hai Vu, Thi Thanh Hai Tran and Thi Lan Le</i>	191
Advertisement Image Classification Using Convolutional Neural Network <i>Tien An Vo, Son Hai Tran and Thai Le</i>	197
Dialog Act Segmentation from Vietnamese Human-Human Conversation <i>Ngo Thi Lan, Pham Khac Linh, Cao Minh-Son, Phan Xuan Hieu and Pham Son Bao</i>	203
Paddy Rice Mapping in Red River Delta region Using Landsat 8 Images: Preliminary results <i>Chuc Man Duc, Anh Nguyen Hoang, Thuy Nguyen Thanh, Hung Bui Quang and Thi Nhat Thanh Nguyen</i>	209
Intent Extraction from Social Media Texts Using Sequential Segmentation and Deep Learning Models <i>Le T. Luong, Son M. Cao, Thang D. Le and Hieu X. Phan</i>	215
A Potential Approach for Emotion Prediction Using Heart Rate Signals <i>Thanh Nam Nguyen, Van Nhan Nguyen, My Huynh Tran Thi and Binh Nguyen</i>	221
The Influence of Knowledge Sharing Behavior and Transactive Memory Systems on Innovative Work Behavior: A Conceptual Model <i>Dong Phung, Igor Hawryszkiewicz and Binh Minh Ha</i>	227
Optimizing Parameters of The Shunt Active Power Filter Using Genetic Algorithm <i>Phan Thanh Hien, Dang Van Huyen, Nguyen Duy An and Nguyen Duy Cuong</i>	233
WikTDV: Data extraction and vector representation resource for Wiktionary senses <i>Danilo Carvalho and Le Minh Nguyen</i>	239
TMACT: A Thematic Map Automatically Creating Tool For Maintaining WebGIS Systems <i>Thang Luu, Thanh Nguyen Thi Nhat, Thuy Nguyen Thanh and Hung Bui Quang</i>	245
A Comprehensive Study on Predicting River Runoff <i>Thi Tran Thi Truong, Hieu Duong Ngoc, Hien Nguyen, Giang Ngo Ngoc Hoang and Nghi Vu Van</i>	251
A Novel Approach Based on Deep Learning Techniques and UAVs to Yield Assessment of Paddy Fields <i>Tri Nguyen Cao, Hieu Duong Ngoc, Van Hoai Tran, Hoa Tran Van, Vu Nguyen and Vaclav Snasel</i>	257
Graph-based visual instance mining with geometric matching and nearest candidates selection <i>Ngoc-Bao Nguyen, Khang M. T. T. Nguyen, Cuong Mai Van and Duy-Dinh Le</i>	263
e-Shoes: Smart Shoes for Unobtrusive Human Activity Recognition <i>Cuong Pham, Diep Nguyen Ngoc and Tu Minh Phuong</i>	269
Designing an annotation scheme for summarizing of Japanese judgment documents <i>Hiroaki Yamada, Simone Teufel and Takenobu Tokunaga</i>	275

Development of Virtual Campus Using 3D GIS Technology: a case study for Vietnam National University, Hanoi <i>Phan Anh, Man Duc Chuc, Bui Quang Hung and Nguyen Thi Nhat Thanh</i>	281
Color anomaly detection on UAV images applicable for search and rescue work <i>Hoai Dao Khanh and Phuong Nguyen Van</i>	287
Improving Chemical-induced Disease Relation Extraction with Learned Features Based on Convolutional Neural Network <i>Hoang-Quynh Le, Duy-Cat Can, Thanh Hai Dang, Mai-Vu Tran, Quang-Thuy Ha and Nigel Collier</i>	292
Merged Grid - An algorithm for continuous constrained k nearest neighbor monitoring <i>Bao L. Nguyen, Tri Q. M. Nguyen and Tien B. Dinh</i>	298
Low Complexity Approaches of K-Best MIMO Decoder for 802.11ac WLAN System <i>Duc Khai Lam, Van Manh Vu and Ngoc Tien Nguyen</i>	304
Author index	309

Examining the Text-length Factor in Evaluating the Readability of Literary Texts in Vietnamese Textbooks

An-Vinh Luong*, Diep Nguyen[†], Dien Dinh[‡]

*Faculty of Information Technology, Ho Chi Minh City University of Science, VNU-HCM, Vietnam
Email: anvinhluong@gmail.com

[†]Department of Linguistics, Ho Chi Minh City University of Social Sciences and Humanities, VNU-HCM, Vietnam
Email: nhudiep2004@gmail.com

[‡]Faculty of Information Technology, Ho Chi Minh City University of Science, VNU-HCM, Vietnam
Email: ddien@fit.hcmus.edu.vn

Abstract—Text readability has an important role in text drafting and document selecting. Researches on the readability of the text have been made long ago for English and some common languages. There are few studies in Vietnamese text readability and most of them are performed from more than two decades ago on very small corpora. Most of the studies in text readability have few mentions of the impact of the text length in evaluating the text. This paper presents our works on examining the role of text length in assessing the readability. The experiment results show that the features related to the text length have a huge impact on Vietnamese text readability assessment for textbooks.

Index Terms—Vietnamese text readability, Text difficulty, Text-length features, School textbooks

I. INTRODUCTION

Text readability — as defined by Edgar Dale and Jeanne Chall [1] — is “the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers has with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting”. Text readability has a huge impact on the reading and comprehending a text. Based on the readability, readers can determine whether a text is suitable for their reading ability or not; the text author(s) can also use the readability of the draft to guide readers or have some adjustments to make it fit toward the reader.

Building a model to analyze text readability has meant a lot in the scientific and practical: help scientists writing research reports more readable; support educators drafting textbooks and curricula to suit each age of students; support publishers in shaping the audience; help governments drafting legal documents to suit the majority of citizens; or to assist manufacturers in preparing user guide for their products *etc.* In addition, text readability can effectively support in choosing appropriate curriculum when teaching language for foreigners.

Researches on text readability have begun since the early years of the 20th century, most of them are for English and some common languages. In Vietnamese, there are only two studies on text readability of Liem Thanh Nguyen and Alan B. Henkin in 1982 and 1985.

Most famous studies in text readability are creating linear functions to assess and grade documents like Dale-Chall formula [1], Flesch Reading Ease formula [2], Flesch-Kincaid formula [3], Gunning Fog formula [4], SMOG formula [5], *etc.*

Recently, with the development of computer, some researchers have applied machine learning for examining text readability:

- In 2001, Si and Callan [6] presented a method of using language model in combination with sentence length to classify the readability of web page. The EM algorithm was used to predict the readability of some web pages in three levels (Kindergarten–Grade2, Grade3–Grade5, and Grade6–Grade8) with about 75% accuracy. In 2005, Collins-Thompson and Callan [7] enhanced Si and Callan’s language model and use Naïve Bayes to predict the level of web page and had a high correlation with labeled grade level.
- In 2005, Schwarm and Ostendorf [8] combined 12 statistical language models, 4 syntactic features and some traditional features like average sentence length, average number of syllables per word, *etc.* to classify the readability levels of the text using SVM classifier. Their results show that syntactic features do not contribute much to the model when examining the Weekly Reader corpus. In 2007, Heilman et al. [9] re-examined such syntactic features in English documents for foreigner and found that they may play more important role in L2 documents readability assessing than in L1 documents.
- In a publication in 2010, Tanaka-Ishii et al. [10] see readability as a sorting problem: instead of creating a text classifier, they created a comparator model to compare pairs of document. The model was trained using some vocabulary features extracted from a corpus with only 2 reading levels — difficult and easy — using SVM classifier. Based on this comparator, they can sort all documents by their text readability with low cost for the training corpus.

TABLE I
STATISTICAL NUMBERS OF EACH GRADE.

	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Over-all
Number of documents	67	62	40	40	28	13	17	21	288
Average number of sentences	17.91	19.11	21.33	20.83	54.25	45.31	63.82	105.62	32.92
Average number of words	158.24	192.61	232.03	244.80	680.89	678.31	971.06	1451.52	404.48
Average number of syllables	178.48	221.98	276.08	287.85	784.11	820.85	1131.47	1709.62	472.36
Average number of characters	826.64	1065.15	1334.25	1395.55	3707.64	3940.92	5400.12	8155.90	2252.56
Average number of distinct words	100.81	125.79	144.63	153.03	305.54	330.62	395.47	527.38	198.30
Average number of distinct syllables	111.36	141.53	164.78	173.23	327.54	372.46	428.35	555.48	217.76
Average sentence length in words	9.40	10.89	11.76	13.03	14.16	18.35	18.26	18.74	12.62
Average sentence length in syllables	10.64	12.65	14.23	15.59	16.27	22.68	21.84	22.57	14.88
Average sentence length in characters	49.55	61.01	69.30	76.25	77.26	110.49	105.82	108.85	71.56
Average word length in syllables	1.13	1.16	1.20	1.18	1.15	1.23	1.19	1.20	1.17
Average word length in characters	5.24	5.59	5.82	5.76	5.41	5.96	5.72	5.76	5.58

TABLE II
STATISTICAL NUMBERS OF EACH GROUP-OF-2-GRADE.

	Grade 2-3	Grade 4-5	Grade 6-7	Grade 8-9	Over-all
Number of documents	129	80	41	38	288
Average number of sentences	18.49	21.08	51.41	86.92	32.92
Average number of words	174.76	238.41	680.07	1236.58	404.48
Average number of syllables	199.39	281.96	795.76	1450.97	472.36
Average number of characters	941.27	1364.90	3781.61	6923.05	2252.56
Average number of distinct words	112.81	148.83	313.49	468.37	198.30
Average number of distinct syllables	125.86	169.00	341.78	498.61	217.76
Average sentence length in words	10.12	12.40	15.48	18.52	12.62
Average sentence length in syllables	11.60	14.91	18.31	22.24	14.88
Average sentence length in characters	55.06	72.77	87.80	107.49	71.56
Average word length in syllables	1.14	1.19	1.17	1.19	1.17
Average word length in characters	5.41	5.79	5.59	5.74	5.58

- In 2012, Vajjala and Meurers [11] examined some Second Language Acquisition (SLA) features in lexical (Lexical Density, Noun Variation, *etc.*) and syntactic level (Mean length of the clauses, Number of Clauses per Sentence, *etc.*) in combination with other traditional features to classify text. The experimental results in a merged corpus collected from WeeklyReader newspaper and BBC-Bitesize website show that assessing text readability English documents for people who English is not the first language needs specific than traditional features.

Although studies on text readability using machine learning had very positive results, many types of features have been examined and experimented with a variety of corpora, but there is a point remain unclear and this study wants to explore: Most of those studies have few mentions of the impact of the text length in assessing text readability. If mentioned, the length of the text is only used to standardize other features such as average sentence and word length, the ratio of difficult words, *etc.* In some case, such as investigating readability for documents in textbooks, the length of the text can be very important.

In this paper, we will examine the impact of the text

length in assessing text readability for Vietnamese school textbooks using SVM classifier. Following this, the rest of this paper is presented as below: we first describe our corpus used for experiments in section II; we then describe some common shallow features used for text readability assessing along with some features related to text-length in section III; after that, section IV describes our experiments and results and our explanations about our result; finally, section V is our discussions and conclusions.

II. CORPUS

We use documents extracted from the textbooks of Vietnamese for elementary students and Literature for middle school students as experimental corpus. In Vietnam, the primary school is divided into five school years — from grade 1 to grade 5. However, Vietnamese language textbooks for grade 1 are only exercises for reading simple letters and words, so we did not collect grade 1 textbooks. At the middle school, the course is divided into four academic years — from grades 6 through 9.

Because there is no digital resource of these textbooks, documents need to be collected manually. First, we scanned

TABLE III
STATISTICAL NUMBERS OF EACH SCHOOL.

	Primary school	Middle school	Over-all
Number of documents	209	79	288
Average number of sentences	19.48	68.49	32.92
Average number of words	199.12	947.76	404.48
Average number of syllables	231.00	1110.92	472.36
Average number of characters	1103.43	5292.68	2252.56
Average number of distinct words	126.60	387.99	198.30
Average number of distinct syllables	142.37	417.22	217.76
Average sentence length in words	10.99	16.95	12.62
Average sentence length in syllables	12.87	20.20	14.88
Average sentence length in characters	61.84	97.27	71.56
Average word length in syllables	1.16	1.18	1.17
Average word length in characters	5.55	5.66	5.58

their hard-copy version into digital images; next, we converted them into text format (using OCR) and then post edited them by the following task:

- 1) Spelling correction: spelling errors are unavoidable in any corpora, especially corpora with a large amount of data. This corpus is not an exception. There are a lot of errors in most of the documents needing to be corrected, especially the documents archived by OCR. With these documents, we had to read all of them to correct their spelling manually without any automation spelling correction tool. As is well-known, most spelling correction tools are based on n-gram for detecting and correcting, so the n-gram will work well if in a sequence of text having few error words. However, documents archived by OCR were full of errors in a sequence, so automation tools were not appropriate in this case.
- 2) Punctuation standardized: in the documents, punctuation marks like dot (.), comma (,), semi-colon (;), colon (:), exclamation (!), question (?), single quotation ('), double quotation ("), brackets ([], (), { }), hyphen (-), slash (/), *etc.* stand close to their related word, we had to separated them by a space (" ") to make the documents clearer, and the statistical operations in these documents more exactly.
- 3) Sentence segmentation and word segmentation: sentences and words are two frequent factors appearing in most researches on readability.

In this research, we used the tool "CLC_VN_Toolkit" for punctuation and encoding standardizing, sentence and word segmentation. This is a tool of COMPUTATIONAL LINGUISTICS CENTER¹, including the functions for the text pre-processing such as sentence segmentation, word segmentation, *etc.*

In this paper, we group collected corpus by 3 levels of grouping for experiment. The first level is by school: by this way, documents are divided into 2 groups: the first is

Primary School group and the second is Middle School group. The second level is by each 2 continuous grade: each group contains two continuous grades from easy to difficult: Group 2–3, Group 4–5, Group 6–7 and Group 8–9. And the last level is by each grade, with a total of 8 groups from grade 2 to grade 9. The detailed statistical numbers of the corpus are presented in Tables I, II and III.

III. FEATURES

In this section, we will describe some shallow features, including common features and text length features used in our experiments.

A. Common features

Average sentence length: the average sentence length of a text is one of the simplest and common characteristic when measuring text readability. In this paper, we examined three types of average sentence length as Equation 1, 2 and 3, where sentence, word, syllable and character count are the total number of sentences, words, syllables and characters in the text:

Average Sentence Length in Words (ASLW):

$$ASLW = \frac{\text{word count}}{\text{sentence count}} \quad (1)$$

Average Sentence Length in Syllables (ASLS):

$$ASLS = \frac{\text{syllable count}}{\text{sentence count}} \quad (2)$$

Average Sentence Length in Characters (ASLC):

$$ASLC = \frac{\text{character count}}{\text{sentence count}} \quad (3)$$

Average word length: two types of average word length were examined in this study as Equation 4 and 5, where word, syllable and character count are the total number of words, syllables and characters in the text:

Average Word Length in Syllables (AWLS):

$$AWLS = \frac{\text{syllable count}}{\text{word count}} \quad (4)$$

¹CLC — <http://www.clc.hcmus.edu.vn> (University of Science, Vietnam National University Ho Chi Minh City)

TABLE IV
CLASSIFICATION RESULTS PERFORMED ON THE GRADE-BY-GRADE DOCUMENTS.

FEATURES	ACCURACY
NoWo, NoCha, ASLC, AWLS, AWLC, PDW, PDS	0.5211
NoSyl, ASLC, AWLS	0.5208
NoSen, NoWo, NoCha, ASLS, AWLC, PDW, PDS	0.5071
NoSen, NoSyl, NoWo, NoDWo, AWLC, PDW, PDS	0.5070
NoSen, NoWo, NoDWo, NoSyl, AWLS, AWLC, PDW, PDS	0.5070
...	...
ASLS, AWLS	0.2569
AWLS, PDW, PDS	0.2326
AWLS, PDW	0.2157
AWLS, PDS	0.2082
PDW, PDS	0.2082

TABLE V
CLASSIFICATION RESULTS PERFORMED ON THE GROUPED-BY-2-CONTINUOUS-GRADES DOCUMENTS.

FEATURES	ACCURACY
NoSen, NoSyl, NoDSyl, AWLS, PDW	0.7536
NoSen, NoSyl, NoDSyl, PDW, PDS	0.7500
NoSen, NoSyl, AWLS, PDW	0.7497
NoSen, NoWo, NoDSyl, AWLS, PDW	0.7497
NoSen, NoSyl, NoDSyl, AWLS	0.7464
...	...
AWLC, PDS	0.4516
AWLS, PDS	0.4481
AWLS, PDW	0.4480
PDW, PDS	0.4480
AWLS, PDW, PDS	0.4478

Average Word Length in Characters (AWLC):

$$AWLC = \frac{\text{character count}}{\text{word count}} \quad (5)$$

In Vietnamese, word maybe monosyllabic or polysyllabic (compound word) with a whitespace between each syllable; each syllable is a combination of letters with or without tonal and word marks. For example: the word “*nghe*” (listen) is a monosyllable with four letters (*n*, *g*, *h*, and *e*); the word “*chay*” (run) is a monosyllable with four letters (*c*, *h*, *a*, *y*) and a tonal mark (*.*); the word “*thời gian*” (time) is a polysyllable (2 syllables “*thời*” and “*gian*”) with eight letters (*t*, *h*, *o*, *i*, *g*, *i*, *a*, *n*), a word mark (‘ ’), a tonal mark (‘ ’) and a whitespace. In our work, each letter, tonal mark, word mark and whitespace was counted as a character.

Percentage of difficult words: in many studies, the percentage of difficult words is an important feature when evaluating text readability. However, creating the easy or difficult word list needs a lot of effort, so most researches used frequent word list as a replacement: if a word does not appear in the frequent list, it will be considered as a difficult word. In this study, we extracted the top 3,000 frequent words from the frequent word list of Dien and Hao [12]. The percentage of difficult words is calculated as Equation 6, where difficult word count are the

total number of difficult words and word count are the total number of words in the text:

$$PDW = \frac{\text{difficult word count}}{\text{word count}} \quad (6)$$

Not only the percentage of difficult words, but also the **percentage of difficult syllables** was examined in our study. We used the top 3,000 frequent syllables extracted from the list of Dien and Hao [12] for experiment. The percentage of difficult syllables is calculated as Equation 7, where difficult syllable count are the total number of difficult syllables and syllable count are the total number of syllables in the text:

$$PDS = \frac{\text{difficult syllable count}}{\text{syllable count}} \quad (7)$$

B. Text length features

Most text readability studies have few mentions of the role of the text length in assessing text readability. If mentioned, the length of the text is only used to standardize other features such as average sentence and word length, the ratio of difficult words, *etc.* In some case, for readability for documents in textbooks, the text length can be very important. If the length is too long or too short, it can affect the reading comprehension in a lesson: Students in higher grades have higher reading

TABLE VI
CLASSIFICATION RESULTS PERFORMED ON THE GROUPED-BY-SCHOOL DOCUMENTS.

FEATURES	ACCURACY
NoSen, NoCha, AWLS, PDW, PDS	0.9620
NoSyl, NoCha, ASLS, ASLC, AWLS	0.9620
NoWo, NoSyl, NoCha, NoDWo, NoDSyl, ASLS, ASLC, AWLC, PDW	0.9619
NoSen, NoCha, NoDWo, ASLW	0.9618
NoSyl, NoCha, ASLC, ASLW, ASLS, AWLS, PDW	0.9617
...	...
AWLS, PDW	0.7256
AWLS, PDS	0.7255
AWLC, AWLS, PDW, PDS	0.7255
AWLC, AWLS, PDS	0.7254
AWLC, PDW, PDS	0.7254

levels and comprehensibility, so they can read longer texts; conversely, students in lower classes are less likely to read and comprehend a long text in the time of a lesson. As we can see in Table I, II and III, the number of words, syllables, characters, *etc.* in documents increases gradually by grade. Therefore, the text length should be a good factor for readability assessment. Some text length features examined in this study are total number of Sentences (**NoSen**), Words (**NoWo**), Syllables (**NoSyl**), Characters (**NoCha**), total number of Distinct Words (**NoDWo**) and total number of Distinct Syllables (**NoDSyl**).

IV. EXPERIMENT AND RESULTS

In this study, we used SVM to classify texts by readability. In this work, the readability of each text is its Grade level or Group-of-2-Grade level. All of the combinations of features mentioned in Section III are used to build classification models to find the best combinations. In order to avoid over-fitting, in each model, the data sets are randomly divided into 5 parts for cross-validation, of which 4 are for the model training and the rest for model accuracy. The experiment was performed in Python using scikit-learn — an open source Python library that implements a range of machine learning, pre-processing, cross-validation and visualization algorithms using a unified interface. Two SVM kernels — Linear and RBF — were used in the experiment. However, the RBF kernel, although has a very fast run rate, has relatively low performance compared to the Linear kernel, so we don't show its results in this paper. Table IV, V and VI show the accuracy of some models, including highest and lowest results.

As Table IV, V and VI show, most of the best classification results have the participation of some text length features such as total number of Sentences (**NoSen**), total number of Words (**NoWo**), total number of Syllables (**NoSyl**), *etc.* Conversely, the worst classification results are almost without the participation of text length features. In the classification of the grade-by-grade documents, the combination (**NoWo**, **NoCha**, **ASLC**, **AWLS**, **AWLC**, **PDW**, **PDS**) had the best result but needs 7 features, the combination (**NoSyl**, **ASLC**, **AWLS**) has an equivalent result but only need a small number of

features (3 features). In the classification of 2-continuous-grades documents, the combination (**NoSen**, **NoSyl**, **AWLS**, **PDW**) yielded high accuracy with only 4 features. Similarly, in the classification of grouped-by-school documents, the combination (**NoSen**, **NoCha**, **NoDWo**, **ASLW**) achieved high accuracy with only 4 features.

We can also see that, if we group the corpus at a more detailed level, the accuracy of the classifiers is also reduced (96% at the field level, 75% at the two-layer level, and 52% for each class). In previous studies, most researchers split the corpus at 2–3 grades a group or the easy-to-medium-difficulty level, rather than dividing into each grade. This is quite consistent with the fact that if a text is written for students of a particular grade, if no knowledge is needed from previous texts, students of the adjacent grades can also read and comprehend that text.

V. CONCLUSION

In this paper, we investigated the impact of the text length in assessing text readability for Vietnamese school textbooks at primary and middle school. The statistical numbers of the collected corpus show that the values of features related to the text length the number of words, syllables, characters, *etc.* in documents increase gradually by grade. Therefore, they should be good features for readability assessment. The classification results using SVM on all features combinations have confirmed this claim.

However, the experiments are only performed on the corpus of Vietnamese documents for elementary students and literature for middle school students, where the text length plays a very important role in ensuring the teaching time: students in higher grades have higher reading levels and comprehensibility, so they can read longer texts; conversely, students in lower classes are less likely to read and comprehend a long text in the time of a lesson. In other domain like common stories or fiction, or specialized textbooks, the text length may be varied because they are not limited by the time of a lesson.

For the future works, we will collect literature document from high school textbooks for experiments. Other deeper

features like part-of-speech, sentence structure, discourse, *etc.* will be examined to create precise classifiers. Some machine learning methods will be examined to create some classifier for automatically Vietnamese text readability assessment.

REFERENCES

- [1] E. Dale and J. S. Chall, "The concept of readability," *Elementary English*, vol. 26, no. 1, pp. 19–26, 1949. [Online]. Available: <http://www.jstor.org/stable/41383594>
- [2] R. Flesch, *The Art of Readable Writing*. New York: Harper and Brothers Publishers, 1949.
- [3] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Technical Training*, vol. Research B, no. February, p. 49, 1975. [Online]. Available: <http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf>
- [4] G. Robert, *The technique of clear writing*. New York: McGraw-Hill Book Co., 1952.
- [5] H. M. Laughlin, "SMOG Grading-a New Readability Formula," *Journal of Reading*, vol. 12, no. 8, p. 639–646, 1969. [Online]. Available: <http://dx.doi.org/10.2307/40011226>
- [6] L. Si and J. Callan, "A statistical model for scientific readability," in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ser. CIKM '01. New York, NY, USA: ACM, 2001, pp. 574–576. [Online]. Available: <http://doi.acm.org/10.1145/502585.502695>
- [7] K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 13, pp. 1448–1462, Nov. 2005. [Online]. Available: <http://dx.doi.org/10.1002/asi.20243>
- [8] S. E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 523–530. [Online]. Available: <https://doi.org/10.3115/1219840.1219905>
- [9] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, April 2007, pp. 460–467. [Online]. Available: <http://www.aclweb.org/anthology/N/N07/N07-1058>
- [10] K. Tanaka-Ishii, S. Tezuka, and H. Terada, "Sorting texts by readability," *Comput. Linguist.*, vol. 36, no. 2, pp. 203–227, Jun. 2010. [Online]. Available: <http://dx.doi.org/10.1162/coli.09-036-R2-08-050>
- [11] S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 163–173. [Online]. Available: <http://www.aclweb.org/anthology/W12-2019>
- [12] D. Đình and D. D. Hao, "Chữ quốc ngữ hiện nay qua các con số thống kê (Current National Vietnamese language through statistics)," in *Hội thảo cấp Quốc gia về chữ quốc ngữ: sự hình thành, phát triển và những đóng góp vào văn hóa Việt Nam (National workshop about National Vietnamese language: the formation, development and contributions to Vietnam culture)*, PhuYen, Vietnam, Oct 2015.