Andrea Tagarelli
Hanghang Tong (Eds.)

LNCS 11917

Computational Data and Social Networks

8th International Conference, CSoNet 2019
Ho Chi Minh City, Vietnam, November 18–20, 2019
Proceedings

Springer

Andrea Tagarelli · Hanghang Tong (Eds.)

Computational Data and Social Networks

8th International Conference, CSoNet 2019
Ho Chi Minh City, Vietnam, November 18–20, 2019
Proceedings

Springer

Editors
Andrea Tagarelli
University of Calabria
Rende, Italy

Hanghang Tong
University of Illinois
at Urbana-Champaign
Urbana, IL, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-34979-0 ISBN 978-3-030-34980-6 (eBook)
https://doi.org/10.1007/978-3-030-34980-6

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Combinatorial Optimization and Learning

Influence Modeling, Propagation, and Maximization

NLP and Affective Computing

Computational Methods for Social Good

User Profiling and Behavior Modeling

Assessing the Readability of Vietnamese Texts Through Comparison
(Extended Abstract)

An-Vinh Luong[1] and Phuoc Tran[2](✉)

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
anvinhluong@gmail.com
[2] NPL-KD Lab, Faculty of Information Technology, Ton Duc Thang University,
Ho Chi Minh City, Vietnam
tranthanhphuoc@tdtu.edu.vn

Abstract. Text readability is a measure of how easy/difficult a text is to read. Studies on readability have been noticed for a long time in English and other popular languages. In Vietnamese, studies on the text readability are still quite limited, mainly because of the lack of corpora with capacity large enough to carry out the examination. In this paper, we present a method for assessing the readability of Vietnamese texts that does not need too much cost to develop a training corpus, based on comparing the correlation difficulty between texts but still achieving positive results.

Keywords: Text readability · Text comparison · Vietnamese language

Text readability—as the definition of Bailin and Grafstein [1]—is a measure of how easy/difficult a text is to read. Studies on readability have been noticed for a long time in English and other popular languages, such as the work of Chen and Meurers [2], Jiang et al. [3]. In Vietnamese, research on the readability of texts is still quite limited, the main reason is the lack of corpora that was graded according to the difficulty level (often with the participation of language experts and at a high cost). In this study, we present a method for assessing the Vietnamese texts readability, using a small number of standard texts (texts that was graded according to the difficulty level), through comparing the correlate readability of the texts.

We collected 100 texts in the field of literature: children's stories, short stories, novels. We randomly selected some texts and gave them to the language experts to read. Each document is given to 2 experts to assess the readability according to 3 levels: Very easy (texts for children at the Elementary School); Easy (texts for people at the Middle School); and Medium (texts for people at the High School). If the evaluation results of both experts are the same, the text will be selected as a standard text. We conduct the such choice and evaluation process until each level has enough 10 texts (the total of 3 levels is 30 documents). These 30 texts are standard texts which will be used to evaluate the remaining texts.

All documents collected are extracted features for training models: Average sentence length calculated on word and syllable; Average word length calculated on syllable; Ratio

© Springer Nature Switzerland AG 2019
A. Tagarelli and H. Tong (Eds.): CSoNet 2019, LNCS 11917, pp. 201–202, 2019.
https://doi.org/10.1007/978-3-030-34980-6_22

of difficult words and Ratio of difficult syllables; Ratio of Sino-Vietnamese words; Ration of local words; Ratio of proper nouns and Ratio of distinct proper nouns. With standard texts, we construct vectors of text pairs by taking all 2-permutations, totally we have 870 text pairs. The feature vectors of each pair is created by vectors subtraction. For example, with 2 texts *a* and *b*, v_a and v_b are 2 feature vectors of *a* and *b*; the feature vector of text pairs *(a, b)*, denoted as v_{ab}; calculated as $v_{ab} = v_a - v_b$. The vector v_{ab} are labeled as follows: if *a* is more difficult than *b* then $v_{ab} = 1$; if *b* is more difficult than *a* then $v_{ab} = -1$; otherwise $v_{ab} = 0$. All these vectors will be used for training a text comparative model using SVM. Using k-fold cross validation, we have checked the accuracy of the model achieved 74.94%.

For the remaining texts, we compare the readability with the standard texts: for each remaining text *x*, we create all feature vectors of text pairs *(x, a)*, with *a* is a text in standard texts, by $v_x - v_a$. These feature vectors will, in turn, be evaluated by the trained comparison model. Finally, we classify the remaining texts through the comparison results: A text *x* will be classified as: (1) Medium if it is more difficult than minimum *k%* Easy and Very easy texts and it is easier than maximum *t%* Medium texts; (2) Very easy if it is easier than minimum *k%* Easy and Medium texts and it is more difficult than maximum *t%* Very easy texts; (3) Easy it is more difficult than minimum *k%* Very easy texts and easier than minimum *k%* Medium texts and it is more difficult than maximum *t%* Easy texts and it is easier than maximum *t%* Easy texts; (4) otherwise, it will be labeled as unclassifiable. *k* and *t* are classification thresholds, obtained through experiments: the higher the value of *k* and *t*, the larger the number of unclassified documents, but the precision of the classification results is higher. In this study, we used *k* = *60%* and *t* = *50%*.

Finally, we selected 30 newly classified documents for re-testing by experts. Each text will be read by an expert to decide if the classification results are proper. The classification results are presented in Table 1. The Precision, Recall and F1 score are 78.13%, 35.71% and 49.02% accordingly.

Table 1. Classification results.

No. of texts	Very easy	Easy	Medium	Unclassifiable
	18	9	5	38

For conclusion, with only a few pre-classified documents, we can build a model to assess the correlate readability between Vietnamese texts with an acceptable accuracy. Based on that assessment model, we have classified the texts by readability with an accuracy of over 78%.

References

1. Bailin, A., Grafstein, A.: Readability: Text and Context. Palgrave Macmillan, UK (2016)
2. Chen, X., Meurers, D.: Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. J. Res. Reading (2018)
3. Jiang, Z., et al.: Enriching word embeddings with domain knowledge for readability assessment. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA (2018)