Readability Classification of Bangla Texts

Zahurul Islam, Md. Rashedur Rahman, and Alexander Mehler

WG Text-Technology
Computer Science
Goethe-University Frankfurt
{zahurul,mehler}@em.uni-frankfurt.de, kamol.sustcse@gmail.com

Abstract. Readability classification is an important application of *Natural Language Processing*. It aims at judging the quality of documents and to assist writers to identify possible problems. This paper presents a readability classifier for Bangla textbooks using information-theoretic and lexical features. All together 18 features are explored to achieve an F-score of 86.46%. The paper is an extension of our previous work [1].

Keywords: Bangla, text readability, information-theoretic features.

1 Introduction

Readability classification aims at measuring how well and easy a text can be read and understood [2]. It deals with mapping texts onto degrees of readability. Thus, readability classification can be reconstructed as a sort of automatic text categorization [3]. Various factors influence the readability of a text including simple features such as type face, font size and text vocabulary as well as more complex features relating to the syntax, semantics, or rhetorical structure of a text [1].

Professionals, such as teachers, journalists, or editors, produce texts for specific audiences. They need to check the readability of their output. Readability classifiers are also used as a means of pre-processing in the framework of *natural language processing* (NLP) [1].

A lot of research on readability classification exists for English [4–9], German [10], French [11], Japanese [12] and Chinese [13]. All these languages are considered as high-resourced languages. They are contrasted with low-resourced languages which are spoken by members of a small community or for which only few resources (corpora, tools etc.) exist [14]. Bangla is a low-resourced language in the latter sense. As an Indo-Aryan language it is spoken in Southeast Asia, specifically in present day Bangladesh and the Indian states of West Bengal, Assam, Tripura and Andaman and on the Nicobar Islands. With nearly 250 million speakers [15], Bangla is spoken by a large speech community. Nevertheless, it is low-resourced because of the lack of appropriate corpora and tools. Thus, though many texts are produced in Bangla everyday, authors can hardly measure their readability due to the lack of appropriate readability classifiers.

Recently, some approaches addressed the readability of Bangla text. Das and Roychudhury [16, 17] experimented with two classical readability measures for English nd applied them to Bangla texts. Sinha et al. [18] proposed two alternative readability measures for Bangla. Islam et al. [1] built a readability classifier using a corpus of Bangla

A. Gelbukh (Ed.): CICLing 2014, Part II, LNCS 8404, pp. 507–518, 2014.

textbooks. Although the classifier achieves an *F-score* of 72.10%, classifiers that produce better F-scores are still required. In this paper, we provide such a better performing readability classifier for Bangla. This is done by example of an extended version of the corpus used in [1]. The corpus is extracted from textbooks used in consecutive grades of the school system of Bangladesh.

Syntactic, semantic and discourse related features are now broadly explored for building readability classifiers for high-resourced languages. Obviously, it is a challenge to do the same for low-resourced languages that lack preprocessing tools. Thus, in this paper, we explore lexical and information-theoretic features which do not require (much) linguistic preprocessing.

The paper is organized as follows: Section 2 discusses related work followed by a description of the underlying corpus (Section 3). The operative readability features are described in Section 4. An experiment based on these features is the topic of Section 5. Its results are discussed in Section 6. Finally, a conclusion is given in Section 7.

2 Related Work

Since the early twentieth century, researchers proposed different readability measures for English [4–9]. All of them explore simple surface-structural features such as *average sentence length* (ASL), *average word length* (AWL) and *average number of syllables in a word*. Many commercial readability tools use these classical measures. Fitzsimmons et al. [19] stated that the SMOG [9] readability measure should be preferred to assess the readability of texts on health care.

Petersen & Ostendorf [20] and Feng et al. [21] show that the classical models have significant drawbacks. Due to recent achievements in linguistic data processing, models of linguistic features are now in the focus of readability studies. [1] summarizes related work regarding language model-based features [22–26], PoS-related features [21, 24, 27, 28], syntactic features [27, 29–32], and semantic features [21, 32].

Recently, Hancke et al. [10] measured the readability of German texts using lexical and syntactic features in conjunction with language models. According to their findings, morphological features influence the readability of German texts. Vajjala and Meurers [33] used lexical features from the field of *Second Language Acquisition* (SLA). In our study, we use *type token ration* (TTR) related readability measures as studied by Vajjala and Meurers [33].

Only few approaches consider the readability of Bangla texts. Das and Roychudhury [16, 17] show that readability measures proposed by Kincaid et al. [7] and Gunning [6] work well for Bangla. However, the measures were tested only for seven documents, mostly novels.

In our previous study [1], we proposed a readability classifier for Bangla using *entropy* and *relative entropy*-based features. We achieved an *F-Score* of 72.10% by combining these features with lexical ones. Recently, Sinha et al. [18] proposed two readability measures that are similar to classical readability measures for English. They conducted a user experiment to identify important structural parameters of Bangla texts.