Text Readability Assessment for Second Language Learners

Menglin Xia and **Ekaterina Kochmar** and **Ted Briscoe**
University of Cambridge
William Gates Building
Cambridge, CB3 0FD, UK
{mx223,ek358,ejb1}@cl.cam.ac.uk

Abstract

This paper addresses the task of readability assessment for the texts aimed at second language (L2) learners. One of the major challenges in this task is the lack of significantly sized level-annotated data. For the present work, we collected a dataset of CEFR-graded texts tailored for learners of English as an L2 and investigated text readability assessment for both native and L2 learners. We applied a generalization method to adapt models trained on larger native corpora to estimate text readability for learners, and explored domain adaptation and self-learning techniques to make use of the native data to improve system performance on the limited L2 data. In our experiments, the best performing model for readability on learner texts achieves an accuracy of 0.797 and PCC of 0.938.

1 Introduction

Developing reading ability is an essential part of language acquisition. However, finding proper reading materials for training language learners at a specific level of proficiency is a demanding and time-consuming task for English instructors as well as the readers themselves. To automate the process of reading material selection and the assessment of reading ability for non-native learners, a system that focuses on text readability analysis for L2 learners can be developed. Such a system enhances many pedagogical applications by supporting readers in their second language education.

Text readability, which has been formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material (Dale and Chall, 1949), is influenced by multiple variables. These may include the style of writing, its format and organization, reader's background and interest as well as various contextual dimensions of the text, such as its lexical and syntactic complexity, level of conceptual familiarity, logical sophistication and so on.

The choice of the criteria to measure readability often depends upon the need and characteristics of the target readers. Most of the studies so far have evaluated text difficulty as judged by native speakers, despite the fact that text comprehensibility can be perceived very differently by L2 learners. In the case of L2 learners, due to the difference in the pace of language acquisition, the focus in readability measures often differs from that for native readers. For example, the grammatical aspects of readability usually contribute more to text comprehensibility for L2 learners than the conceptual cognition difficulty of the reading material (Heilman et al., 2007). A system that is tailored towards learner's perception of reading difficulty can produce more accurate estimation of text reading difficulty for non-native readers and thus better facilitate language learning.

One of the major challenges for a data-driven approach to text readability assessment for L2 learners is that there is not enough significantly sized, properly annotated data for this task. At the same time, text readability assessment in general has been previously studied by many researchers and there are a number of existing corpora aimed at native speakers that can be used. To address the problem, we compiled a collection of texts that are tailored for

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 12–22,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

L2 learners' readability and looked at several approaches to make use of existing native data to estimate readability for L2 learners.

In sum, the contribution of our work is threefold. First, we develop a system that produces state-of-the-art estimation of text readability, exploit a range of readability measures and investigate their predictive power. Second, we focus on readability for L2 learners of English and present a level-graded dataset for non-native readability analysis. Third, we explore methods that help to make use of the existing native corpora to produce better estimation of readability when there is not enough data aimed at L2 learners. Specifically, we apply a generalization method to adapt models trained on native data to estimate text readability for learners, and explore domain adaptation and self-training techniques to improve system performance on the data aimed at L2 learners. To the best of our knowledge, these approaches have not been applied in readability experiments before. The best performing model in our experiments achieves an accuracy (ACC) of 0.797 and Pearson correlation coefficient (PCC) of 0.938.

2 Related Work

2.1 Automated Readability Assessment

Many previous studies on text readability assessment have used machine learning based approaches, which enable investigation of a broader set of linguistic features. Si and Callan (2001) and Collins-Thompson and Callan (2004) were among the early works on statistical readability assessment. They applied unigram language models and naïve Bayes classification to estimate the grade level of a given text. Experiments showed that the language modelling approach yields better results in terms of accuracy than the traditional readability formulae, such as the the Flesch-Kincaid score (Kincaid et al., 1975). Schwarm and Ostendorf (2005) extended this method to multiple language models. They combined traditional reading metrics with statistical language models as well as some basic parse tree features and then applied an SVM classifier. Heilman et al. (2007; 2008) expanded the feature set to include certain lexical and grammatical features extracted from parse trees while using a linear regression model to predict the grade level.

Pitler and Nenkova (2008) and Feng et al. (2010) were the first to introduce discourse-based features into the framework. The experiments with discourse features demonstrated promising results in predicting the readability level of text for both classification and regression approaches.

Kate et al. (2010) looked at both the effect of the feature choice and the machine learning framework choice on performance, and found that the improvement resulting from changing the framework is smaller than that from changing the features.

2.2 Readability Assessment for L2 Learners

Most previous work on readability assessment is directed at predicting reading difficulty for native readers. Several efforts in developing automated readability assessment that take L2 learners into consideration have emerged since 2007. Heilman et al. (2007) tested the effect of grammatical features for both L1 (first language) and L2 readers and found that grammatical features play a more important role in L2 readability prediction than in L1 readability prediction. Vajjala and Meurers (2012) combined measures from Second Language Acquisition research with traditional readability features and showed that the use of lexical and syntactic features for measuring language development of L2 learners has a substantial positive impact on readability classification. They observed that lexical features perform better than syntactic features, and that the traditional features have a good predictive power when used with other features. Shen et al. (2013) developed a language-independent approach to automatic text difficulty assessment for L2 learners. They treated the task of reading level assessment as a discriminative problem and applied a regression approach using a set of features that they claim to be language-independent. However, most of these studies have used textual data annotated with the readability levels for native speakers of English rather than L2 learners specifically.

While the majority of work on automated readability assessment are for English, studies on L2 readability in other languages, including French (François and Fairon, 2012), Portuguese (Branco et al., 2014), and Swedish (Pilán et al., 2015), are also emerging. These studies generally use textbook materials with readability levels assigned by publishers

	Level1	Level2	Level3	Level4	Level5
age group	7-8	8-9	9-10	10-14	14-16
original corpus	629	801	814	1969	3500
modified corpus	529	767	801	1288	845

Table 1: Number of documents in the original and modified WeeBit corpus

Exams	KET	PET	FCE	CAE	CPE
targeted level	A2	B1	B2	C1	C2
# of docs	64	60	71	67	69
avg. len. of text	14.75	19.48	38.07	45.76	39.97

Table 2: Statistics for the Cambridge English Exams data

or language instructors.

Overall, study of automatic readability analysis for L2 learners is still in its early stages, mainly due to the lack of available well-labelled data annotated with the readability levels for L2 learners.

3 Data

3.1 Native Data: the WeeBit Corpus

Among the existing publicly available corpora, the WeeBit corpus created by Vajjala and Meurers (2012) is one of the largest datasets for readability analysis. The WeeBit corpus is composed of articles targeted at readers of different age groups from two sources, the Weekly Reader magazine and the BBC-Bitesize website. Within the dataset, the Weekly Reader data consists of texts covering age-appropriate non-fictional content for four grade levels, corresponding to children of ages between 7-8, 8-9, 9-10 and 10-12 years old. The BBC-Bitesize website data is targeted at two grade levels, for ages between 11-14 and 14-16. The two datasets are merged to form the WeeBit corpus, with the targeted ages used to assign readability levels.

A copy of the original WeeBit corpus was obtained from the authors (Vajjala and Meurers, 2012). The texts are webpage documents stored in raw HTML format. We have identified that some texts contain broken sentences or extraneous content from the webpages, such as copyright declaration and links, that correlate with the target labels in a way which is likely to artificially boost performance on the task and would not generalize well to other datasets. To avoid that, we re-extracted texts from the raw HTML and discarded text documents that do not contain proper reading passages. Table 1 shows the distribution of texts in the modified dataset.

3.2 L2 Data: the Cambridge Exams dataset

Most work on readability assessment has been done on native corpora with age-specific reading levels (Schwarm and Ostendorf, 2005; Feng et al., 2010). Such texts are aimed not at L2 learners but rather at native-speaking children of different ages. Therefore, the level annotation in such texts is arrived at using criteria different from those that are relevant for L2 readers. The lack of significantly sized L2 level-annotated data raises a problem for readability analysis aimed at L2 readers. To tackle this, we created a dataset with texts tailored for L2 learners' readability specifically.

We have collected a dataset composed of reading passages from the five main suite Cambridge English Exams (KET, PET, FCE, CAE, CPE).[1] These five exams are targeted at learners at A2–C2 levels of the Common European Framework of Reference (CEFR) (Council of Europe, 2001).[2] The documents are harvested from all the tasks in the past reading papers for each of the exams. The Cambridge English Exams are designed for L2 learners specifically and the A2–C2 levels assigned to each reading paper can be treated as the level of reading difficulty of the documents for the L2 learners.[3] Table 2 shows the number of documents at each CEFR level across the dataset. The data is available at `http://www.cl.cam.ac.uk/~mx223/cedata.html`.

Experimenting on the language testing data annotated with the L2 learner readability levels is one of the contributions of this research. Most previous work on readability assessment for English have relied on the data annotated with readability levels aimed at native speakers. In this work, we use language testing data with the levels assigned based on L2 learner levels, and we believe that this level annotation is more appropriate for text readability assessment for L2 learners than using texts with the level annotation aimed at native speakers.

[1] `http://www.cambridgeenglish.org`

[2] The CEFR determines foreign language proficiency at six levels in increasing order: A1 and A2, B1 and B2, C1 and C2.

[3] We are aware that the type of the task may also have an effect on the reading difficulty of the texts, but this is ignored at this stage.