CEFR-Based Sentence Difficulty Annotation and Assessment

Yuki Arase[†] and **Satoru Uchida**[⋆] and **Tomoyuki Kajiwara**[◇]
[†]Graduate School of Information Science and Technology, Osaka University, Japan
[⋆]Faculty of Languages and Cultures, Kyushu University, Japan
[◇]Graduate School of Science and Engineering, Ehime University, Japan
arase@ist.osaka-u.ac.jp, uchida@flc.kyushu-u.ac.jp
kajiwara@cs.ehime-u.ac.jp

Abstract

Controllable text simplification is a crucial assistive technique for language learning and teaching. One of the primary factors hindering its advancement is the lack of a corpus annotated with sentence difficulty levels based on language ability descriptions. To address this problem, we created the CEFR-based Sentence Profile (CEFR-SP) corpus, containing 17k English sentences annotated with the levels based on the Common European Framework of Reference for Languages assigned by English-education professionals. In addition, we propose a sentence-level assessment model to handle unbalanced level distribution because the most basic and highly proficient sentences are naturally scarce. In the experiments in this study, our method achieved a macro-F1 score of 84.5% in the level assessment, thus outperforming strong baselines employed in readability assessment.

1 Introduction

Controllable text simplification, first proposed by Scarton and Specia (2018), is the automatic rewriting of sentences to make them comprehensible to a target audience with a specific proficiency level. Among its primary applications are providing reading assistance to language learners and helping teachers adjust the difficulty level of their teaching materials (Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014; Paetzold, 2016). The fine-grained control of output levels to match the linguistic ability of the readership is crucial for these educational applications.

While readability assessments have been actively studied (*e.g.*, in (Vajjala Balakrishna, 2015; Meng et al., 2020; Deutsch et al., 2020)), linking readability to language ability is difficult. Readability scores, such as the Flesch–Kincaid grade level (Kincaid et al., 1975), are intended for native speakers, not for language learners to whom very different considerations apply. Pilán et al. (2014) and Ozasa et al. (2007) revealed that readability metrics designed for L1 do not apply to L2 learners. Furthermore, readability definitions use documents rather than sentences, which are required by text simplification at the sentence-level, as their unit.

The lack of a corpus annotated by sentence difficulty level hinders the advancement of controllable text simplification. Previous studies (Scarton and Specia, 2018; Nishihara et al., 2019; Agrawal et al., 2021) necessarily used corpora annotated for readability rather than difficulty; furthermore, they assumed that all sentences in a document had the same readability (*i.e.*, the document level in Newsela (Xu et al., 2015)).

To solve these problems, we created a large-scale English corpus annotated by sentence difficulty levels based on the Common European Framework of Reference for Languages (CEFR),[1] the most widely used international standard describing learners' language ability. Our CEFR-based Sentence Profile (CEFR-SP) corpus adapts CEFR to sentence levels. A sentence is categorised as a certain level if a person with the corresponding CEFR-level can readily understand it. CEFR-SP provides CEFR levels for 17k sentences annotated by professionals with rich experience teaching English in higher education.

A major challenge in sentence-level assessment is the unbalanced distribution of levels: sentences at the basic (A1) and highly proficient (C2) levels are naturally scarce. To handle this, we propose a sentence-level assessment model with a macro-F1 score of 84.5%. We designed a metric-based classification method with a simple inductive bias that avoids overfitting to majority classes (Vinyals et al., 2016; Snell et al., 2017). Our method generates embeddings representing each CEFR-level and estimates a sentence's level based on its cosine similarity to these embeddings. Empirical results confirm that our method effectively copes with unbalanced

[1] https://www.coe.int/en/web/common-european-framework-reference-languages

label distribution and outperforms the strong baselines employed in readability assessments.

This study makes two main contributions. First, we present the largest corpus to date of sentences annotated according to established language ability indicators. Second, we propose a sentence-level assessment model to handle unbalanced label distribution. CEFR-SP and sentence-level assessment codes are available[2] for future research at `https://github.com/yukiar/CEFR-SP`.

2 Related Work

Related studies have assessed text levels on different granularity (document and sentence) and level definitions (readability/complexity and CEFR).

2.1 Document-based Readability

Previous studies have assessed readability and created corpora with document readability annotations. WeeBit (Vajjala and Meurers, 2012), the OneStopEnglish corpus (Vajjala and Lučić, 2018), and Newsela provide manually written documents for various readability levels. Working with these annotated corpora, previous studies have used various linguistic and psycholinguistic features to develop models for assessing document-based readability (Heilman et al., 2007; Kate et al., 2010; Vajjala and Meurers, 2012; Xia et al., 2016; Vajjala and Lučić, 2018). Neural network-based approaches have proven to be better than feature-based models (Azpiazu and Pera, 2019; Meng et al., 2020; Imperial, 2021; Martinc et al., 2021). In particular, Deutsch et al. (2020) showed that pre-trained language models outperform feature-based approaches, and the combination of linguistic features plays no role in performance gains.

2.2 Sentence-based Readability

Previous studies annotated sentences' *complexities* based on crowd workers' subjective perceptions. Stajner et al. (2017) used a 5-point scale to rate the complexity of sentences written by humans or generated by text simplification models. Brunato et al. (2018) used a 7-point scale for sentences extracted from the news sections of treebanks (McDonald et al., 2013). However, as Section 3.4 confirms, relating complexity to language ability descriptions is challenging. Naderi et al. (2019) annotated German sentence complexity based on language learners' subjective judgements. In contrast, the CEFR-level of a sentence should be judged *objectively* based on the understanding of language learners' skills. Hence, we presume that a sentence CEFR-level can be judged only by language education professionals based on their teaching experience. For sentence-based readability assessments, previous studies regarded all sentences in a document to have the same readability (Collins-Thompson and Callan, 2004; Dell'Orletta et al., 2011; Vajjala and Meurers, 2014; Ambati et al., 2016; Howcroft and Demberg, 2017). As we show in Section 3.4, this assumption hardly holds.

The *simplicity* of a sentence is one of the primary aspects in a text simplification evaluation, which is commonly judged by human. There are a few corpora annotated by the sentence simplicity for automatic quality estimation of text simplification (Štajner et al., 2016; Alva-Manchego et al., 2021). Nakamachi et al. (2020) applied a pretrained language model for estimating the sentence simplicity and used it to reward a reinforcement learning–based text simplification model. The sentence simplicity is distinctive from CEFR levels based on the established language ability descriptions.

2.3 CEFR-based Text Levels

Attempts have been made to establish criteria for CEFR-level assessments. For example, the English Profile (Salamoura and Saville, 2010) and CEFR-J (Ishii and Tono, 2018) projects relate English vocabulary and grammar to CEFR levels based on learner-written' and textbook corpora. Tools such as Text Inspector[3] and CVLA (Uchida and Negishi, 2018) endeavour to measure the level of English reading passages automatically. Xia et al. (2016) collected reading passages from Cambridge English Exams and predicted their CEFR levels using features proposed to assess readability. Rama and Vajjala (2021) demonstrated that Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) consistently achieved high accuracy for multilingual CEFR-level classification.

Although these micro- (*i.e.*, vocabulary and grammar) and macro-level (*i.e.*, passage-level) approaches have proven useful, few attempts have been made to assign CEFR levels at the *sentence* level, despite its importance in learning and teaching. Pilán et al. (2014) conducted a sentence-level assessment for Swedish based on CEFR; however,

[2] The licenses of the data sources are detailed in Ethics Statement section.

[3] `https://textinspector.com/`

they regarded document-based levels as sentence levels. Furthermore, their level assessment was as coarse as predicting either above B1 or not.

3 CEFR-SP Corpus

This section describes the design of the annotation procedure and discusses sentence-level profiles. CEFR describes language ability on a 6-point scale: A1 indicates the proficiency of beginners; A2, B1, B2, C1, and C2 indicates mastery of a language at the basic (A), independent (B), and proficient (C) levels. Because CEFR is skill-based, each level is defined by 'can-do' descriptors indicating what learners can do,[4] CEFR levels for sentences cannot be defined directly.

Therefore, we used a bottom-up approach, assigning CEFR levels to sentences based on the 'can-do' descriptors of reading skills under the definition that a sentence is, for example, at A1 level if it can be readily understood by A1-level learners. We hypothesise that with sufficient teaching experience and CEFR knowledge, it is possible to objectively determine at which level a learner can understand each sentence. We therefore carefully selected annotators with sufficient expertise through pilot and trial sessions.

3.1 Annotation Procedure

Pilot Study A pilot study was conducted to verify the hypothesis that sufficient teaching experience and CEFR knowledge will allow an objective evaluation of sentence levels. We recruited participants with three levels of expertise to label 228 sample sentences: an English-language education specialist with 12 years of teaching experience in higher education, a graduate student majoring in English education who is familiar with CEFR, and a group of three graduate students with various majors (natural language processing and ornithology) and no prior knowledge of CEFR or English-teaching experience. The results showed that the second expert had a high agreement rate with the first senior expert (Pearson correlation coefficient 0.74), whereas the members of the third group agreed less often with the senior expert (Pearson correlation coefficients: 0.45, 0.50, and 0.59). These results confirm that annotators with considerable experience and knowledge agree on the judgement of the CEFR levels of sentences.

Annotation Guidelines The annotators were familiarised with the annotation guidelines before beginning their work. The guidelines described the scales and 'can-do' descriptions of CEFR reading skills with example sentences of each level that were assessed by the expert. Importantly, the guidelines required the annotators to judge each sentence's level based on their English-teaching experience. Annotators were allowed to look in a dictionary to establish word levels but were instructed not to determine a sentence's level solely based on the levels of the words it contained.

Annotator Selection For formal annotation, we recruited eight annotators with diversified English-teaching experience. We then conducted a trial session in which the annotators were asked to label 100 samples extracted from the target corpora of formal annotation. These samples were labelled by the senior expert in the pilot study as references. Pearson correlation coefficients against the expert ranged from 0.59 to 0.77, roughly correlating with the participants' experience in English-teaching in terms of duration (years of teaching) and role (as private tutor or teacher in higher education). We finally selected two having high agreement rates (Pearson correlation coefficients: 0.75 and 0.73) and small average level-assignment differences (0.11 and 0.22) compared to the expert.[5] The annotation guidelines were finalised to provide example sentences with corresponding CEFR levels on which multiple annotators had agreed in the pilot and trial sessions.

3.2 Sentence Selection

Sentences were drawn from Newsela-Auto, Wiki-Auto, and the Sentence Corpus of Remedial English (SCoRE). Newsela-Auto and Wiki-Auto, created by Jiang et al. (2020), are specifically used for text simplification.[6] SCoRE (Chujo et al., 2015) was created for computer-assisted English learning, particularly for second language learners with lower-level proficiency. The sentences in SCoRE were carefully written by native English speakers, understanding the educational goals of each proficiency level; they include A-level sentences, which are scarce in text simplification corpora.

[4] https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52

[5] CEFR levels were converted into a 6-point scale.

[6] With the plan of expanding CEFR-SP to a parallel corpus in the future, we included parallel sentences. Note that our data-split policy (Section 5.1) ensures that highly similar sentences do NOT appear in training and validation/test sets.