	A1	A2	B1	B2	C1	C2
Train	535	3,646	8,996	6,636	1,908	100
Validation	125	568	1,130	821	290	74
Test	111	561	1,148	826	292	74

Table 5: Distribution of sentence levels in training, validation, and test sets

Because a level is allowed to have multiple prototypes, an initialisation vector is generated for the k-th prototype at level i, $\hat{\boldsymbol{c}}_i^k \in \hat{\boldsymbol{C}}$, by adding Gaussian noise with mean $\mu = 0$ and variance σ^2 set to 5% of that computed on all elements in $\hat{\boldsymbol{c}}_0, \hat{\boldsymbol{c}}_1, \ldots, \hat{\boldsymbol{c}}_{J-1}$:

$$\hat{\boldsymbol{c}}_i^k = \hat{\boldsymbol{c}}_i + \mathcal{N}(\mu, \sigma^2).$$

Finally, expecting these prototypes to capture the distinctive features of different levels, we orthogonalise the matrix $\hat{\boldsymbol{C}}$ and set the initial values of the prototype matrix $\boldsymbol{C}$.

5 Evaluation

In this section, the proposed level assessment model is evaluated using the CEFR-SP corpus.

5.1 Corpus Splitting

We split CEFR-SP into three sets: approximately 80% for training, 10% for validation, and 10% for the test set, as shown in Table 5. We adjusted the number of sentences for infrequent levels to preserve a reasonable number of test and validation cases.[13] In corpus splitting, we ensured that highly similar sentences did not appear in both the training and validation/test sets, as detailed in Appendix B.

A sentence in CEFR-SP may have as many as two levels, both assignments being regarded as equally reliable. Therefore, the predictions during training, validating, and testing were assumed correct if they matched either of the annotated labels.

5.2 Evaluation Metrics

The ability to predict *all* levels correctly is important for educational applications. As the distribution of levels was unbalanced, the models were evaluated using macro-F1 to penalise models that ignored minor classes. In addition, because CEFR levels are ordinal, the models were also evaluated using the quadratic weighted kappa (Bakeman and Gottman, 1997).

To reduce the dependence of performance fluctuation on initialisation seeds, the experiments were conducted 12 times with randomly selected seeds. We then discarded the best and worst results and reported a mean macro-F1 score and kappa value with a 95% confidence interval.

5.3 Setting

We used BERT-Base, cased model (Devlin et al., 2019) as the pretrained MLM to encode sentences in the models that were compared.[14] Specifically, we used the outputs of the 11-th layer, which performed strongly. K, the number of prototypes of the proposed method, was set to 3 to maximise the average macro-F1 of the validation set in the 1–10 range.

Comparison Because of the roughly positive correlation between the word and sentence levels (Section 3.3), we implemented a bag-of-words (BoW)[15] classifier using support vector machines (Cortes and Vapnik, 1995) as the naive baseline. Moreover, as a simpler variant of metric-based classification method, we implemented a k-nearest neighbour (kNN) (Fix and Hodges, 1989) classifier. We used mean-pooled token embeddings of freezed BERT as features and the cosine distance for distance computation. The size of k was set to 6 which marked the highest macro-F1 on the development set.

As the state-of-the-art baseline, we used a BERT-based classifier that outperforms conventional linguistic-feature-based classifiers in predicting passage-level readability (Deutsch et al., 2020) and CEFR levels (Rama and Vajjala, 2021), as well as on simple and complex binary classification (Garbacea et al., 2021) of the WikiLarge corpus (Zhang and Lapata, 2017). The proposed model was compared with these baselines with or without loss weighting.

Ablation Study We investigated the effect of K with an ablation study. We also implemented variations of the proposed method without loss weighting and initialisation based on sentence embed-

[13] We tentatively used the higher level among the two annotated labels for assigning a sentence into either the training, validation, or test sets.

[14] In a preliminary experiment, we compared BERT, RoBERTa (Liu et al., 2019), and Sentence-BERT (Reimers and Gurevych, 2019) with different configurations and confirmed that there was no significant difference between them. Therefore, we decided to use the standard BERT-Base.

[15] Word-level features performed much worse and were omitted in this experiment.

		A1	A2	B1	B2	C1	C2	Average	Weighted κ
BoW	w/o lossW	0.0	69.7	76.3	66.4	34.7	0.0	41.2	0.354 ± 0.000
		44.2	64.9	73.0	69.6	53.8	8.0	52.3	0.429 ± 0.000
kNN		1.5 ± 1.4	75.2 ± 0.7	81.8 ± 0.4	66.4 ± 0.6	8.1 ± 2.6	0.0 ± 0.0	38.8 ± 0.4	0.373 ± 0.004
BERT	w/o lossW	12.8 ± 9.4	$\mathbf{83.6 \pm 0.3}$	87.0 ± 1.1	$\mathbf{86.7 \pm 1.2}$	82.9 ± 1.5	76.8 ± 5.5	71.7 ± 1.7	0.592 ± 0.012
		72.7 ± 3.9	82.7 ± 1.1	85.5 ± 0.9	86.4 ± 0.7	84.9 ± 1.2	83.6 ± 3.0	82.5 ± 0.9	0.609 ± 0.014
Proposed	w/o lossW	12.0 ± 13.4	$\mathbf{83.6 \pm 0.4}$	$\mathbf{87.8 \pm 1.2}$	86.3 ± 1.4	83.0 ± 0.9	0.0 ± 0.0	58.7 ± 1.8	0.595 ± 0.013
	w/o init	76.1 ± 1.5	80.5 ± 1.4	84.7 ± 1.4	85.7 ± 1.3	85.3 ± 1.2	88.1 ± 2.1	83.3 ± 0.9	$\mathbf{0.628 \pm 0.017}$
		$\mathbf{78.0 \pm 1.3}$	81.4 ± 0.9	86.5 ± 1.1	85.9 ± 0.8	$\mathbf{85.4 \pm 1.3}$	$\mathbf{89.7 \pm 1.6}$	$\mathbf{84.5 \pm 0.7}$	$\mathbf{0.628 \pm 0.010}$

Table 6: Macro-F1 scores (%) per level and quadratic weighted kappa values measured on the CEFR-SP test set; 'w/o lossW' indicates a model without loss weights and 'w/o init' indicates a model without initialisation using sentence embeddings. The proposed method (last row) preserves high F1 scores at the infrequent A1 and C2 levels and the best quadratic weighted kappa value.

dings. The former method achieved its maximum validation macro-F1 score when $K = 1$. The latter method used the same settings as the proposed method, except for prototype initialisation; it initialised the prototype embeddings using a normal distribution $\mathcal{N}(0, 1)$.

5.4 Implementation Details

The classifier layer of the BERT baselines comprised a linear layer with weights $W \in \mathbb{R}^{d \times J}$ and a 10% dropout to the input sentence embedding. Other conditions remained the same as those of the proposed method. We input a sentence embedding computed by Equation (1) and calculated the standard classification loss of cross-entropy. Loss weights were computed by Equation (2).

All models were implemented using the PyTorch, Lightning, Transformers (Wolf et al., 2020), and scikit-learn libraries.[16] The neural network models were trained on an NVIDIA Tesla V100 GPU using an AdamW (Loshchilov and Hutter, 2019) optimiser with a batch size of 128. The training was stopped early, with 10 patience epochs and a minimum delta of $1.0e-5$ based on the average macro-F1 score of all levels measured on the validation set. The loss weighting factor α and other hyperparameters were tuned using Optuna (Akiba et al., 2019). For the proposed method and BoW and BERT baselines, α values were set to 0.2, 0.3, and 0.4, respectively. The complete hyperparameter settings are described in Appendix C.

5.5 Results

Table 6 shows the CEFR-SP test set results by means of macro-F1 scores (%) per level and quadratic weighted kappa values with 95% confidence intervals. As in previous studies, the BERT-based classifiers outperformed the BoW baselines. This result confirms that words and their levels, despite their importance, are not solely responsible for determining sentence levels. The kNN classifier showed higher macro-F1 scores than BoW without loss weighting on A2 and B1 because of the powerful BERT embeddings. However, it failed to identify A1 and C levels, which indicates the significance of addressing unbalanced label distribution.

The proposed method (last row) had the highest F1 scores for infrequent levels, *i.e.*, A1 and C2, but a slightly reduced performance for the more common levels. We consider this acceptable, considering the method's capability to assess infrequent levels. Overall, the proposed method achieved the highest average macro-F1 score (84.5%) and quadratic weighted kappa value (0.628).

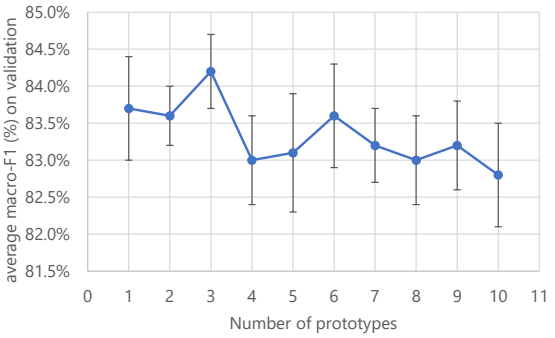

Figure 2: Effects of number of prototypes: average macro-F1 scores (%) measured on validation set

Effects of Loss Weighting While loss weighting is highly effective in alleviating the effects of unbalanced label distribution on all models, it is more critical for the proposed method. Exclusion of loss weighting overlooks the A1 and C2 levels, as is clear from the sixth row of Table 6. Confusion matrices confirmed that A1 and C2 sentences were

[16] Hyperlinks to these libraries are listed in Appendix D.

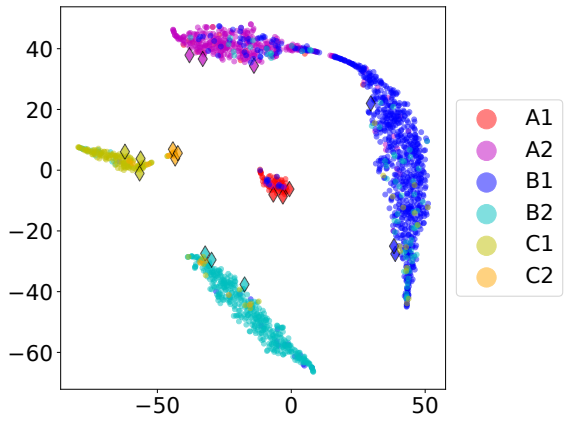

Figure 3: Visualisation of prototypes (represented by ♢) and sentence embeddings of the proposed method.

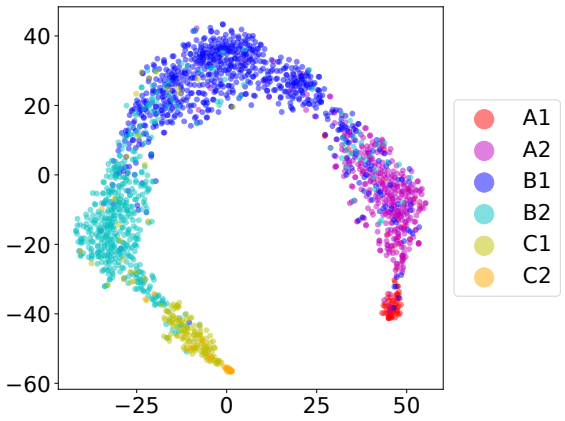

Figure 4: Visualisation of sentence embeddings of BERT baseline with loss weighting.

misclassified to their adjacent levels.

Effects of Initialisation The seventh row of Table 6 presents the results for the proposed method without initialisation using sentence embeddings. This method tended to have larger confidence intervals than the proposed model. Moreover, we observed that it fell into an undesired solution that overlooked A1 and C2 levels depending on initialisation seeds, as reflected in lower macro-F1 scores. These results confirm that our initialisation was effective for training stabilisation.

Effects of Number of Prototypes Figure 2 shows the average macro-F1 scores with 95% confidence intervals measured on the validation set when the number of prototypes in the proposed method changed from 1 to 10. The average macro-F1 score initially improved as the number of prototypes increased; it peaked at three, and then gradually decreased. This trend empirically confirms the effectiveness of multiple prototypes and shows that a relatively small number of prototypes is sufficient for CEFR-SP.

Visualisation Figure 3 plots the sentence embeddings generated by the proposed method, and Figure 4 those generated by the BERT baseline with loss weighting. The gold levels are colour-coded; for the proposed method, the prototypes are indicated by diamond markers. We used T-SNE (van der Maaten and Hinton, 2008) for visualisation, setting the perplexity to 30 and number of iterations to 5k to ensure convergence.

The class boundaries were not clear in the embeddings of the baseline. In contrast, the embeddings of the proposed method formed clear clusters by level owing to the metric-based classification; this improved the interpretability. When assessing the level of a new sentence, the cosine similarity to each prototype indicates whether the assessment result is high-confidence, *i.e.*, prototypes of a single level exhibit significantly high cosine similarity to the sentence, or ambiguous, *i.e.*, multiple levels exhibit competitive cosine similarities.

6 Summary and Future Work

In this study, we introduced CEFR-SP, the first English sentence corpus annotated with CEFR levels. The carefully designed annotation procedure involved recruiting experts with strong backgrounds in English education to ensure the reliability of the assigned labels. CEFR-SP allows the development of an automatic sentence-level assessment model. The proposed method can handle unbalanced level distributions using a metric-based classification.

Our future work will involve collecting parallel sentences of CEFR-SP to make it directly applicable for training text simplification models. We will also develop controllable text simplification models based on reinforcement learning: the proposed level assessment model will be employed to reward the generation of lower-level sentences.

Limitations

Because of severe space constraints, we have reported only the lexical profile of CEFR-SP. We will present its syntactic and psycholinguistic features and analyse it from an educational perspective in a future publication. Moreover, CEFR-SP is not directly applicable to train controllable text simplification models that require parallel sentences