Improving Readability Assessment with Ordinal Log-Loss

Lim, Ho Hung; Lee, John S. Y.

Published in:
Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)

Published: 01/06/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Publication details:
Lim, H. H., & Lee, J. S. Y. (2024). Improving Readability Assessment with Ordinal Log-Loss. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 343–350). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.28

Improving Readability Assessment with Ordinal Log-Loss

Ho Hung Lim and **John S. Y. Lee**
Department of Linguistics and Translation
City University of Hong Kong
limhhresearch@gmail.com,jsylee@cityu.edu.hk

Abstract

Automatic Readability Assessment (ARA) aims to predict the level of difficulty of a text, e.g. at Grade 1 to Grade 12. It can be helpful for teachers and students in identifying and revising text to the desirable level of difficulty. ARA is an ordinal classification task since the predicted levels follow an underlying order, from easy to difficult. However, most neural ARA models ignore the distance between the gold level and predicted level, treating all levels as independent labels. This paper investigates whether distance-sensitive loss functions can improve ARA performance. We evaluate a variety of loss functions on neural ARA models, and show that ordinal log-loss can produce statistically significant improvement over the standard cross-entropy loss in terms of adjacent accuracy in a majority of our datasets.

1 Introduction

Automatic Readability Assessment (ARA) aims to predict the level of difficulty of a text, e.g. at Grade 1 to Grade 12. It can be helpful for teachers and students in identifying and revising text to the desirable level of difficulty. ARA is an ordinal classification task since the levels follow an underlying order, from easy to difficult. Yet, in ARA models trained with traditional machine learning, the use of ordinal classification has yielded mixed results (Heilman et al., 2008; Feng et al., 2010; Jiang et al., 2014). Further, most neural ARA models treat the task as multi-class classification (Xia et al., 2016; Azpiazu and Pera, 2019; Filighera et al., 2019; Tseng et al., 2019; Deutsch et al., 2020; Martinc et al., 2021; Lee et al., 2021) and ignore the distance between the gold level and predicted level. In these models, a classifier is typically trained with the standard cross-entropy loss function, which treats the difficulty levels as independent labels. Further, performance evaluation often penalizes incorrect predictions equally, regardless of their distance from the gold.

Recognizing the ordinal nature of ARA could potentially enhance performance and enable more accurate evaluation. A loss function that reflects label distance could be suitable, since the boundary between difficulty levels may not be clear-cut, especially on fine-grained scales. While severe mistakes are never desirable, a sufficiently close prediction may be acceptable in some applications, such as retrieval of extra-curricular reading materials. Evaluation metrics that reflect the average distance from the gold label would therefore be more informative.

Distance-sensitive loss functions have received relatively little attention in neural ARA. Zeng et al. (2022) showed that soft labels could improve performance, but the evaluation was limited to BERT and only one loss function. We present a more comprehensive study on a variety of loss functions, evaluated on a range of pre-trained language models, hyper-parameters, and performance metrics. Experimental results show that ordinal log-loss (Castagnos et al., 2022) performs best overall for neural ARA models. It achieves a statistically significant improvement over the standard cross-entropy loss in terms of adjacent accuracy in a majority of our datasets, though sometimes at the expense of accuracy.

The rest of the paper is organized as follows. After a review of the major loss functions in Section 2, we give details on the experimental set-up in Section 3. We then report results in Section 4.[1]

2 Previous work

Many text classification tasks, ranging from ARA and essay scoring, to sentiment and review rating prediction, have an ordinal structure. Let $\mathcal{Y} = \{r_1, r_2, ..., r_K\}$ be the set of possible labels. Ordinal binary classification exploits the structure

[1]Code and data can be accessed at https://github.com/hhlim333/Readability-Assessment-with-Ordinal-Log-Loss

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 343–350

with $K-1$ binary classifiers (Frank and Hall, 2001). Ordinal Multi-class Classification with Voting was found to be potentially helpful in improving ARA performance (Jiang et al., 2014). Ordinal regression models have been applied to ARA models trained in traditional machine learning. While Heilman et al. (2008) found that the Proportional Odds Model offered competitive performance, Feng et al. (2010) reported that ordinal classifiers did not perform better than standard classifiers. Loss-sensitive classification, which is the focus of this paper, utilizes loss functions that impose higher penalty to predictions further from the gold label, based on a distance function $d(r_i, r_j)$ that specifies the distance between labels r_i and r_j. Two main families of these loss functions are as follows.

2.1 Soft labels

Soft labels for ordinal regression (Bertinetto et al., 2020) is a distance-sensitive loss function that has been found to be effective for ARA. The soft label is defined as follows:

$$y_i = \frac{\exp\left(-\beta \cdot d(r_i, r_t)\right)}{\sum_{k=1}^{K} \exp\left(-\beta \cdot d(r_k, r_t)\right)} \quad (1)$$

where r_t is the gold label; $r_i \in \mathcal{Y}$ is the i-th label; and the hyperparameter β specifies how much more probability mass to assign to labels closer to the gold.

Zeng et al. (2022) applied the soft label version of Diaz and Marathe (2019) to ARA using a simple distance function: the distance between the gold and an adjacent label is a positive constant, and infinity for all other labels. A BERT-based neural classifier trained on this loss function outperformed the standard cross-entropy loss on both English and Chinese data.

2.2 Ordinal log-loss

Ordinal log-loss (OLL) is defined as:

$$-\sum_{i=1}^{N} \log(1 - p_i) d(y, i)^{\alpha} \quad (2)$$

where the hyperparameter α adjusts the amount of penalty, with a higher value leading to the greater penalty for predicted labels at a longer distance from the gold (Castagnos et al., 2022). OLL is distinguished in its use of the weight $-log(1-p_i)$, rather than p_i as in many other loss functions, to impose greater penalty on more severe errors.

Castagnos et al. (2022) have shown OLL to be beneficial in a number of text classification tasks, but their evaluation focused only on BERT-tiny. This paper is the first attempt to apply OLL on ARA. We conduct a comprehensive study utilizing a variety of loss functions and pre-trained language models, and analyzing trade-off between accuracy and adjacent accuracy.

3 Experimental Set-up

This section describes the loss functions (Section 3.1), the datasets (Section 3.2) and training procedure (Section 3.3).

3.1 Loss functions

We investigate the following loss functions for training neural ARA models:[2]

Baseline The standard cross-entropy loss.

WKL Weighted Kappa Loss (de la Torre et al., 2018).

EMD Earth Mover's Distance (Hou et al., 2016).

OLL-α Ordinal log-loss (Castagnos et al., 2022) with the hyperparameter α, as defined in Section 2.2.

SOFT-β Soft labels (Bertinetto et al., 2020) with the hyperparameter β, as defined in Section 2.1.

Zeng et al The model proposed by Zeng et al. (2022) (Section 2.1), based on the soft label version of Diaz and Marathe (2019), which does not use the β hyperparameter.

Following Castagnos et al. (2022), we tuned the α parameter for OLL on $\{1, 1.5, 2\}$ and the β parameter for SOFT on $\{2, 3, 4\}$. They were optimized on the validation set of the Cambridge Dataset to $\alpha = 1$ and $\beta = 2$, respectively. We used the default distance function $d(r_i, r_j) = |r_i - r_j|$ in all experiments.

3.2 Datasets

Our experiments make use of three English and two Chinese datasets (see detailed statistics in Appendix A):

[2] https://github.com/glanceable-io/ordinal-log-loss

Loss function	Cam		CC		OSE		CMT		CMER	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Baseline	0.387	0.533	1.047	1.729	**0.042**	0.077	1.244	3.524	1.696	5.666
Zeng et al	0.347	0.413	0.953	1.494	0.056	0.084	1.118	2.985	1.681	5.623
OLL-1	0.347	**0.400**	**0.776**	**1.012**	0.074	0.13	**1.112**	**2.894**	**1.638**	**4.847**
SOFT-2	**0.333**	**0.400**	1.035	1.694	**0.042**	0.077	1.159	3.169	1.679	5.592
EMD	0.433	0.553	0.906	1.541	0.046	**0.06**	1.171	3.104	1.664	5.205
WKL	0.867	1.493	1.235	2.671	0.446	0.614	2.252	10.107	3.455	19.177

Table 1: Mean Absolute Error (MAE) and Mean Squared Error (MSE) in ARA using RoBERTa on the English datasets Cambridge (Cam), Common Core (CC) and OneStopEnglish (OSE); and using MacBERT on the Chinese datasets CMT and CMER

Cambridge (Cam) This dataset contains articles for various Cambridge English Exams, labeled with five levels (A2-C2) in the Common European Framework of Reference (Xia et al., 2016). We use the train/validation/test set of the downsampled version provided by Lee et al. (2021), which consists of 60 items per level.[3]

OneStopEnglish (OSE) This corpus consists of 189 aligned texts, each written at three reading levels: beginner, intermediate, and advanced (Vajjala and Lučić, 2018), hence a total of 567 texts.[4]

Common Core (CC) The Common Core corpus consists of 168 texts, labeled at five grade bands (Grades 2–3, 4–5, 6–8, 9-10, and 11–12) from Appendix B of the English Language Arts Standards of the Common Core State Standards (Chen and Meurers, 2016).[5]

China Mainland Textbook (CMT) This corpus consists of a total of 2,723,430 characters, distributed in 2,621 texts in twelve grades, all taken from Chinese textbooks from the first grade of primary school to the third grade of high school in mainland China (Cheng et al., 2019).

China Mainland Extracurricular Reading (CMER) This corpus consists of 2,260 texts distributed at Grade 1 to 12, taken from extracurricular reading books for children and teenagers.[6]

[3]Accessed at https://github.com/brucewlee/
[4]Accessed at https://github.com/nishkalavallabhi/
[5]https://xiaobin.ch/Chen_Meurers_16Frequency/
[6]https://github.com/JinshanZeng/DTRA-Readability

3.3 Pre-trained language models

We evaluated the pre-trained language models BERT, RoBERTa, BART, and XLNET[7] in English experiments. In the Chinese experiments, we used MacBERT[8], which was shown to perform best in previous research on Chinese ARA (Lim et al., 2022). All models were downloaded from HuggingFace transformers v4.5.0 (Wolf et al., 2020).[9]

4 Experimental results

All results are averaged based on stratified 5-fold cross-validation with a 8:1:1 split for train/validation/test. We first report overall results based on Mean Absolute Error (MAE) and Mean Squared Error (MSE) (Section 4.1), and then analyze their performance in terms of adjacent accuracy and accuracy.[10] Henceforth, all Chinese results are based on MacBERT, and the English results on RoBERTa, sicne they performed best among the four PLMs evaluated (see Table 7 in Appendix D).

4.1 Mean Error

Table 1 shows the performance of neural ARA models in terms of MAE and MSE when trained with the loss functions described in Section 3.1. Weighted Kappa Loss (WKL) produced the worst performance, below the standard cross-entropy baseline in all datasets. Earth Mover's Distance (EMD) outperformed the baseline in four out of

[7]https://huggingface.co/bert-base-uncased,roberta-base,bart-base,xlnet-base-cased
[8]https://huggingface.co/hfl/chinese-macbert-large
[9]We used AdamW (optimizer) (Kingma and Ba, 2015), linear (scheduler), 10% (warmup steps), 8 (batch size), 3 (epoch) for all pre-trained language models. For English experiments, we use the learning rate of 2e-5 for BERT and 3e-5 for the other pre-trained language models. For Chinese experiments,we use the learning rate of 2e-5 for MacBERT.
[10]All metrics are calculated with SciKit-learn (Pedregosa et al., 2011).

Loss Function	(a) Accuracy					(b) Adjacent Accuracy				
	Cam	CC	OSE	CMT	CMER	Cam	CC	OSE	CMT	CMER
Baseline	0.68	0.294	0.975	0.364	0.285	0.940	0.659	0.982	0.686	0.561
Zeng et al	0.68	0.318	0.958	0.382	0.277	0.98	0.729	0.986	0.735	0.575
OLL-1	0.673	0.341	0.954	0.368	0.232	**0.987***	**0.882****	0.972	**0.740***	0.563
OLL-1.5	0.64	0.329	0.846	0.316	0.209	0.973	**0.882****	**0.993**	0.738*	0.573
OLL-2	0.56	0.341	0.891	0.317	0.201	0.98	0.824**	0.989	0.731*	**0.583**
SOFT-2	0.693	0.294	0.975	0.381	0.277	0.98	0.671	0.982	0.718*	0.574
SOFT-3	**0.727**	0.294	**0.979**	**0.387**	0.281	0.967	0.682	0.986	0.726*	0.555
SOFT-4	0.713	0.294	**0.979**	0.367	**0.29**	0.96	0.659	0.982	0.699	0.568
EMD	0.62	**0.376**	0.961	0.359	0.243	0.953	0.753	0.993	0.709	0.573
WKL	0.387	0.271	0.639	0.182	0.105	0.8	0.659	0.916	0.5	0.307

Table 2: ARA performance based on (a) accuracy; and (b) adjacent accuracy (* means a statistically significant improvement at $p < 0.05$ according to McNemar's Test over the baseline; ** means statistically significant improvement over both the baseline and the Zeng et al. model)

five datasets, yielding the lowest MSE on OSE. The Zeng et al model improved upon the baseline in all datasets except OSE. SOFT-2 outperformed Zeng et al in three out of the four datasets, and produced the best performance on Cambridge (tied with OLL-1), suggesting the utility of the β hyperparameter. Overall, OLL-1 achieved the best performance, with the smallest MSE on four of the five datasets. In the remainder of the discussion, we will focus on Zeng et al, SOFT-β and OLL-α.

4.2 Adjacent accuracy

Table 2(b) shows the results in terms of adjacent accuracy. The OLL-α models outperformed the baseline in the vast majority of settings, suggesting their ability to reduce severe ARA errors.[11] Of the four PLMs, the best performance was obtained with RoBERTa (Appendix D).

OLL-1 achieved the best adjacent accuracy at 0.987 on Cambridge and 0.882 on Common Core.[12] It also scored the highest Macro F1 and Weighed F1 on these two datasets (see Table 5 in Appendix C). OSE is particularly challenging since the baseline already achieved excellent performance at 0.989 adjacent accuracy; OLL was able to make an improvement on adjacent accuracy and F1 only when α is set to 1.5. OLL-1 improved upon the baseline on both Chinese datasets, and outperformed Zeng et al on CMT.

[11]Among all combinations of α values, PLMs and datasets, there are only two exceptions: OLL-1 with RoBERTa on OSE, and with BART on Cambridge.

[12]Statistically significant at $p = 0.0391$ and $p = 0.0000$, respectively, according to McNemar's Test.

4.3 Accuracy

OLL-1 generally performed worse than the baseline, both in terms of accuracy (Table 2(a))[13] and F1 (Table 6 in Appendix C). SOFT-2 improved upon the baseline and Zeng et al in most settings, although the improvement was not statistically significant.

SOFT-3 established a new state-of-the-art in accuracy and F1 for neural ARA models, on both the Cambridge and OSE datasets. Its performance (accuracy at 0.727 and 0.979, respectively) surpassed the previous best (0.680 and 0.975) in neural models (Lee et al., 2021), although it is still outperformed by hybrid models, which require handcrafted linguistic features. SOFT-3 also obtained the best result in Chinese on CMT (0.387), outperforming the baseline and the Zeng et al model.

5 Conclusion

Since ARA is an ordinal classification task, the magnitude of classification error should in principle be taken into account. This paper has presented a comprehensive evaluation of a variety of loss functions that are sensitive to the distance between the predicted label and gold label.

Our experiments on neural ARA models suggest that ordinal log-loss (OLL) is able to capture the ordinal nature of the task, reducing the mean absolute error and mean squared error on most datasets. It produces significant improvement over the standard cross-entropy function in terms of adjacent accuracy, but at the expense of accuracy in some

[13]We obtained slightly higher accuracy for the baseline on the OSE dataset than reported by Lee et al. (2021).

settings. These results suggest that future ARA models should consider using OLL for applications that need to avoid severe errors but do not require precise classification.

A number of research directions may be pursued. First, ARA accuracy could be further improved by optimizing the distance function in the ordinal log-loss and soft label models. Second, the usability of the ARA model in an educational setting, for example assisting teachers and students in text selection and revision, is also worth investigating.

Acknowledgements

This work was partly supported by the Language Fund from the Standing Committee on Language Education and Research (project EDB(LE)/P&R/EL/203/14) and by a Teaching Development Grant from City University of Hong Kong (project 6000834).

References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. 2020. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

François Castagnos, Martin Mihelich, and Charles Dognin. 2022. A Simple Log-based Loss Function for Ordinal Text Classification. In *Proc. 29th International Conference on Computational Linguistics (COLING)*, pages 4604–4609.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing Text Difficulty with Word Frequencies. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, page 84–94.

Yong Cheng, Dekuan Xu, and Xueqiang Lv. 2019. Automatically Grading Text Difficulty with Multiple Features. *Data Analysis and Knowledge Discovery*, 3(7):103–112.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 4738–4747.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noemie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Proc. COLING*.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, page 335–348. Springer.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *Proc. 12th European Conference on Machine Learning (ECML)*, page 145–156.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proc. Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. In *arXiv preprint arXiv:1611.05916*.

Zhiwei Jiang, Gang Sun, Qing Gu, and Daoxu Chen. 2014. An Ordinal Multi-class Classification Method for Readability Assessment of Chinese Documents. *LNAI*, 8793:61–72.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. 3rd International Conference for Learning Representations*, San Diego.

Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Ho Hung Lim, Tianyuan Cai, John S. Y. Lee, and Meichun Liu. 2022. Robustness of Hybrid Models in Cross-domain Readability Assessment. In *Proc. 20th Workshop of the Australasian Language Technology Association (ALTA)*.

Matej Martinc, Senja Pollak, Marko, and Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. An Innovative BERT-Based Readability Model. In *Lecture Notes in Computer Science, vol 11937*.

Grade	Cam		CC		OSE	
	Texts	Text length	Texts	Text length	Texts	Text length
1	60	140.12	20	294.65	189	531.97
2	60	271.25	30	320.70	189	677.90
3	60	614.50	45	472.09	189	820.76
4	60	778.73	36	549.83	na	na
5	60	761.85	37	612.05	na	na

Table 3: Number of texts and average length at each grade in the Cam, CC and OSE dataset

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*, page 12–22.

Jinshan Zeng, Yudong Xie, Xianglong Yu, John S. Y. Lee, and Ding-Xuan Zhou. 2022. Enhancing Automatic Readability Assessment with Pre-training and Soft Labels for Ordinal Regression. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4586–4597.

A Appendix: Dataset statistics

This section provides detailed statistics for all datasets.

B Appendix: Computing details

We used a NVIDIA Tesla V100 GPU to train 80% of the full dataset. The following is the total training time of the experiments on OLL-1, measured in seconds:

English Experiments (BERT, RoBERTa, XLNet, BART):

- Cambridge (638,496,1410,607)
- OneStopEnglish (1261, 948, 2286, 1110)
- CommonCore (382,310,776,377)

Chinese Experiment (MacBERT):

Grade	CMT		CMER	
	Texts	Text length	Texts	Text length
1	235	108.95	218	145.53
2	320	198.58	217	308.44
3	386	329.48	234	538.35
4	321	425.39	229	628.08
5	282	569.82	200	682.41
6	252	660.89	255	701.29
7	199	1202.13	221	1227.19
8	142	1176.94	205	1324.25
9	134	1443.84	188	1302.54
10	140	1617.08	100	2182.08
11	89	1900.85	96	2252.34
12	121	1930.74	97	2043.69

Table 4: Number of texts and average length at each grade in the CMT and CMER dataset

- CMT (12498)
- CMER (11809)

C Appendix: F1 Evaluation

This section reports F1 evaluation, based on adjacent accuracy (Table 5) and accuracy (Table 6), respectively. We used RoBERTa on the English datasets Cambridge (Cam), Common Core (CC) and OneStopEnglish (OSE); and MacBERT on the Chinese datasets CMT and CMER.

D Appendix: Evaluation on other PLMs

This appendix provides detailed results for all pretrained language models (BERT, RoBERTa, XLNet, BART).

Loss	Macro F1					Weighted F1				
Function	Cam	CC	OSE	CMT	CMER	Cam	CC	OSE	CMT	CMER
Baseline	0.938	0.527	0.982	0.593	0.518	0.938	0.551	0.982	0.655	0.548
Zeng et al	0.98	0.615	0.986	0.647	0.532	0.98	0.658	0.986	0.715	0.562
OLL-1	**0.987**	**0.839**	0.972	0.642	0.502	**0.987**	**0.862**	0.972	0.721	0.544
OLL-1.5	0.973	0.833	**0.993**	**0.661**	0.513	0.973	0.859	**0.993**	**0.722**	0.554
OLL-2	0.98	0.742	0.989	0.631	0.518	0.98	0.788	0.989	0.712	0.563
SOFT-2	0.98	0.544	0.982	0.629	**0.533**	0.98	0.572	0.982	0.694	0.562
SOFT-3	0.966	0.557	0.986	0.636	0.511	0.966	0.588	0.986	0.705	0.539
SOFT-4	0.959	0.527	0.982	0.605	0.528	0.959	0.551	0.982	0.673	0.555
EMD	0.952	0.722	**0.993**	0.62	0.53	0.952	0.722	**0.993**	0.686	**0.565**
WKL	0.766	0.626	0.894	0.437	0.216	0.766	0.605	0.894	0.473	0.248

Table 5: ARA performance in F1, based on *adjacent accuracy*

Loss	Macro F1					Weighted F1				
Function	Cam	CC	OSE	CMT	CMER	Cam	CC	OSE	CMT	CMER
Baseline	0.658	0.091	0.975	0.282	0.253	0.658	0.134	0.975	0.324	0.27
Zeng et al	0.668	0.131	0.958	**0.322**	0.246	0.668	0.173	0.958	**0.363**	0.262
OLL-1	0.654	0.206	0.954	0.279	0.201	0.654	0.242	0.954	0.346	0.221
OLL-1.5	0.591	0.189	0.812	0.236	0.167	0.591	0.226	0.812	0.286	0.181
OLL-2	0.496	0.182	0.868	0.227	0.153	0.496	0.224	0.868	0.273	0.168
SOFT-2	0.68	0.095	0.975	0.305	0.246	0.68	0.139	0.975	0.351	0.262
SOFT-3	**0.717**	0.093	**0.979**	0.318	0.253	**0.717**	0.136	**0.979**	0.361	0.264
SOFT-4	0.699	0.091	**0.979**	0.294	**0.259**	0.699	0.134	**0.979**	0.337	**0.274**
EMD	0.569	**0.243**	0.961	0.288	0.204	0.569	**0.284**	0.961	0.329	0.214
WKL	0.237	0.157	0.539	0.083	0.024	0.237	0.139	0.539	0.08	0.026

Table 6: ARA performance in F1, based on *accuracy*

Metric →		Accuracy			Adjacent Accuracy		
PLM	Loss Func.	Cam	CC	OSE	Cam	CC	OSE
BERT	Baseline	0.573	0.388	**0.919**	0.907	0.694	0.989
	Zeng et al	0.567	**0.4**	0.719	0.94	**0.835**	0.993
	OLL-1	0.5	0.365	0.709	**0.973***	0.812*	0.989
	OLL-1.5	0.44	0.365	0.737	**0.973***	0.788*	**0.996**
	OLL-2	0.467	0.353	0.705	**0.973***	0.765	0.993
	SOFT-2	**0.593**	0.388	0.768	0.913	0.753	0.993
	SOFT-3	0.573	0.376	0.765	0.913	0.718	0.993
	SOFT-4	0.587	0.353	0.754	0.92	0.659	0.993
	WKL	0.407	0.318	0.505	0.813	0.753	0.863
	EMD	0.48	0.353	0.786	0.92	0.776	0.993
RoBERTa	Baseline	0.68	0.294	0.975	0.94	0.659	0.982
	Zeng et al	0.68	0.318	0.958	0.98	0.729	0.986
	OLL-1	0.673	**0.341**	0.954	**0.987***	**0.882****	0.972
	OLL-1.5	0.64	0.329	0.846	0.973	**0.882****	**0.993**
	OLL-2	0.56	**0.341**	0.891	0.98	0.824**	0.989
	SOFT-2	0.693	0.294	0.975	0.98	0.671	0.982
	SOFT-3	**0.727**	0.294	**0.979**	0.967	0.682	0.986
	SOFT-4	0.713	0.294	**0.979**	0.96	0.659	0.982
	WKL	0.387	0.271	0.639	0.8	0.659	0.916
	EMD	0.62	0.376	0.961	0.953	0.753	0.993
BART	Baseline	0.62	0.388	**0.968**	0.927	0.776	0.989
	Zeng et al	0.593	**0.435**	0.944	0.92	0.788	**0.996**
	OLL-1	0.52	0.353	0.965	0.92	0.847	0.993
	OLL-1.5	0.493	0.318	0.958	**0.94**	0.871**	0.993
	OLL-2	0.42	0.294	0.916	**0.94**	**0.882****	**0.996**
	SOFT-2	0.6	0.412	0.947	0.92	0.776	0.993
	SOFT-3	0.6	**0.435**	0.944	0.9	0.8	0.986
	SOFT-4	**0.627**	0.388	0.954	0.907	0.776	0.989
	WKL	0.393	0.294	0.596	0.8	0.612	0.902
	EMD	0.56	0.4	0.961	0.913	0.788	0.993
XLNET	Baseline	0.573	0.365	0.804	0.933	0.671	0.993
	Zeng et al	**0.713**	0.388	0.811	0.933	0.8	0.993
	OLL-1	0.653	0.318	0.737	0.967	0.824*	**0.996**
	OLL-1.5	0.593	0.365	**0.818**	**0.973****	**0.847***	0.993
	OLL-2	0.467	0.329	0.807	**0.973****	0.835*	0.993
	SOFT-2	0.667	0.388	0.877	0.933	0.753	0.993
	SOFT-3	0.653	**0.424**	0.891	0.92	0.741	0.996
	SOFT-4	0.633	0.341	0.853	0.933	0.753	0.993
	WKL	0.42	0.318	0.481	0.86	0.659	0.86
	EMD	0.587	0.318	0.856	0.9	0.741	0.989

Table 7: ARA performance on the English datasets (* means statistically significant improvement at $p < 0.05$ according to McNemar's Test over the baseline; ** means statistically significant improvement over both baseline and Zeng et al.)