PopEval: A Character-Level Approach to End-To-End Evaluation Compatible with Word-Level Benchmark Dataset

Hong-Seok Lee, Youngmin Yoon, Jang Pil Hoon, Chankyu Choi

NAVER Corporation
$\{hongs.lee \cdot youngmin.yoon \cdot ph.jang \cdot chankyu.choi\}@navercorp.com$

***Abstract*—The most prevalent scope of interest for OCR applications used to be scanned documents, but it has now shifted towards the natural scene. Despite the change of times, the existing evaluation methods are still based on the old criteria suited better for the past interests. In this paper, we propose PopEval, a novel evaluation approach for the recent OCR interests. The new and past evaluation algorithms were compared through the results on various datasets and OCR models. Compared to the other evaluation methods, the proposed evaluation algorithm was closer to the human's qualitative evaluation than other existing methods. Although the evaluation algorithm was devised as a character-level approach, the comparative experiment revealed that PopEval is also compatible on existing benchmark datasets annotated at word-level. The proposed evaluation algorithm is not only applicable to current end-to-end tasks, but also suggests a new direction to redesign the evaluation concept for further OCR researches.**

***Keywords*-end-to-end evaluation; character level evaluation; character-oriented evaluation, optical character recognition;**

I. Introduction

Evaluation metric is not merely a performance evaluation and ranking system of competition, but a navigator of optical character recognition(OCR) research because the direction of developing a certain model is heavily affected by its evaluation method. Therefore, the evaluation metric should reflect actual performance of the model. In this study, we investigated the theoretical bases of existing evaluation algorithms, and suggest a novel evaluation concept optimized to tasks of which current OCR researches are mainly focused: robust reading.

In summary, our contributions are as follows: 1) We propose a novel character-oriented end-to-end evaluation protocol, compatible with existing benchmark datasets annotated at word level. 2) To confirm the compatibility between PopEval method and the word-level annotated benchmark datasets, we newly reannotated and published the most widely used test datasets for end-to-end system: focused scene text(ICDAR2013) and incidental scene text(ICDAR2015) at character-level as quadrilaterals [1], [2]. 3) we performed the comparative analysis among evaluation metrics, detection-recognition algorithms and representative test datasets, then the results were compared with human qualitative end-to-end evaluation.

The source code of the PopEval and the test datasets of ICDAR2013 and ICDAR2015 which were newly annotated at character-level are available at: https://github.com/naver/popeval

II. Related works

A. Detection Evaluation

In ICDAR2013 competition, DetEval was adopted as a detector evaluation metric at object level, that determines the matching objects by using double threshold system based on pixel precision and pixel recall [3]. DetEval also handles one-to-many(split) and many-to-one(merge) matching problems, but as it uniformly handles these match cases as same weight irrespective of the match condition, it results in errors. In addition, there have been similarity measuring methods to solve one-to-many and many-to-one problems, but these approaches required feature extraction [4].

ICDAR2015 competition adopted the intersection over union(IOU) based PASCAL EVAL as an evaluation metric [5]. If the IOU between two object areas exceeds 0.5, then the objects are considered as a match. In the IOU method, because a ground truth(GT) object only matches one predicted object, the split and merge problems are ignored [2].

COCO-Text competition adopted average precision(AP) with IOU [6]. It required additional confidence rate values of detected objects to be calculated. The split and merge problems are not handled because it uses the concept of IOU.

B. Recognition Evaluation

For recognition tasks, total edit distance and correctly recognized words rate were adopted as the evaluation metric [1], [2]. The above performance indicator values have been calculated for both case sensitive and case insensitive. In correctly recognized words rate, one exactly matched recognized sequence is counted as one matching case regardless of the length of the transcript.

C. End-to-End Evaluation

Conventional end-to-end evaluation method is a pipeline that consists of detection and recognition phases.

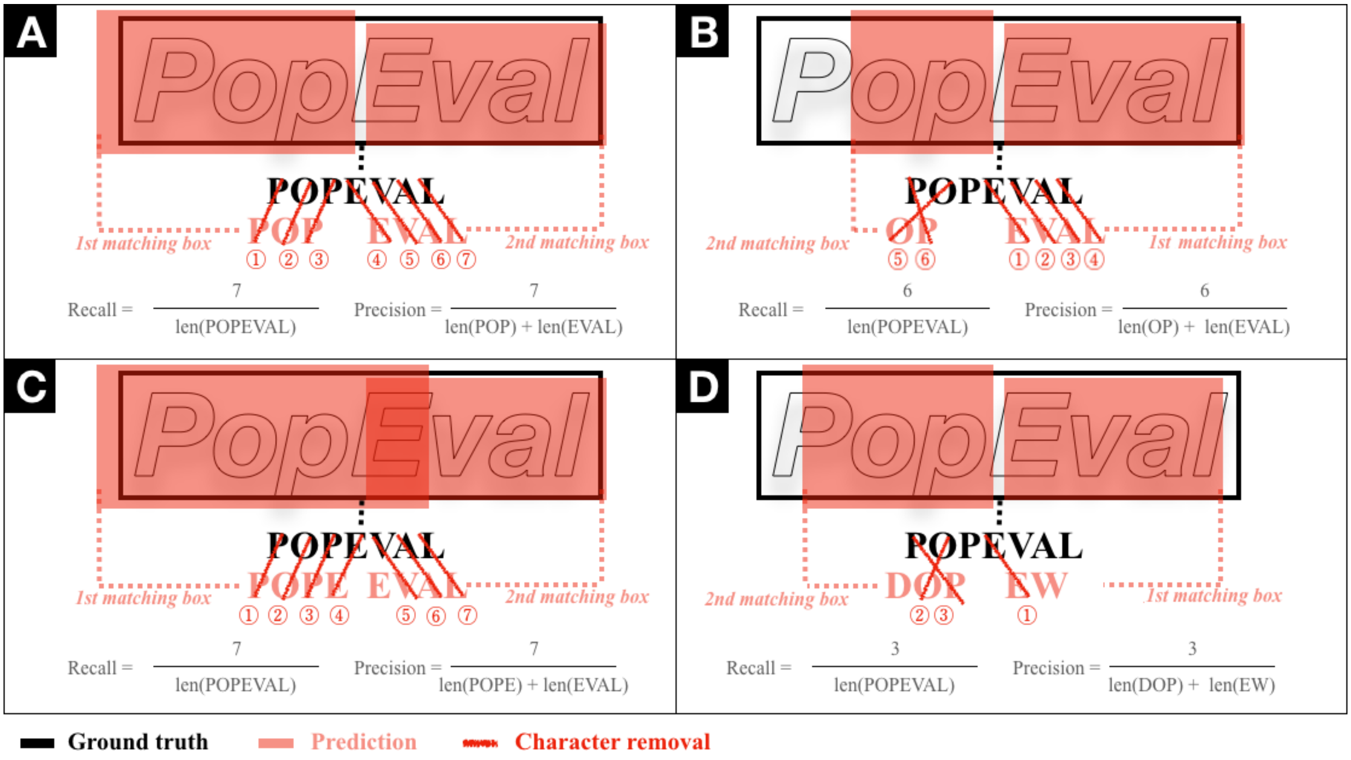

Figure 1. PopEval character removal process scheme. The black box and text are GT and the red box and text are prediction. The order of character removal is indicated as circle character. (A) The GT polygon and text label 'POPEVAL' and The predicted polygons and text labels 'POP', and 'EVAL'; (B) Deletion case: 'OP' and 'EVAL' were predicted. There were six removed characters, one remaining character of GT; (C) Insertion case: 'POPE' and 'EVAL' were predicted. There were seven removed characters, and one remaining character of prediction; (D) Complicated case: 'DOP' and 'EW' were predicted. There were three removed characters, four remaining characters of GT, two remaining character of prediction;

1) Detection: In ICDAR2013, ICDAR2015, and COCO-Text competition, the detected objects of which IOU is greater than 0.5 with the corresponding GT object passes to recognition phase.

2) Recognition: In the ICDAR2013 and ICDAR2015 competitions, the recognition part adopted controlled vocabulary system which defines minimal common conditions to grant meaningful performance comparison.

COCO-Text competition does not use the vocabulary system, and it adopts the case-insensitive correctly recognized words method. By using the correctly recognized words method, the results passed through the detection phase are finally selected, then the corresponding APs are calculated using residues through the both phase [6]. The AP has a drawback that is not intuitive to understand the absolute performance level of a certain model [7].

Otherwise, the normalized edit distance(1-NED) can be used. After the detection phase, the edit distance of the GT and the recognized transcript is divided into the most long length of the GT or the recognized transcripts then subtract the average results to 1. If the detected box is not caught in the detection phase, the calculation of 1-NED is performed by assuming that the recognized transcripts were blank [8].

III. PopEval: Our approach

PopEval is a character level end-to-end evaluation metric that is based on removing overlapping characters between the GT and the OCR result. In this criteria, the number of removed characters is considered as true positive count to be used for calculating character recall and precision. Contrary to existing end-to-end evaluation metrics, PopEval is not just a pipeline structure which consists of detection and recognition parts, but a seamless structure of which character removal is conducted by integration of detection and recognition results.

The principle of PopEval algorithm is to adopt how human beings evaluate a certain recognized result comparing with GT. Based on the principle, there are three criteria in which characters are removed.

A. Determining the Iteration Order of GT Polygons

People read English text as images, from top left to bottom right. Likewise, the iteration order of scattered GT on each image is determined as the distance between the GT polygon centroid and left-top point of the image. The shorter the distance, the earlier the iteration sequence order. This approach solves many-to-one problems in relation to

Algorithm 1: PopEval evaluation metric, part 1 of 2

```
global removed_char_count = 0
def main(GTs, Dets):
    one-to-one, one-to-many = InspectRelation(GTs, Dets)
    # one-to-one = [[GT1, Det1], [GT2, Det2], ...]
    # one-to-many = [[GT3, [Det3, Det4]], ...]
    if len(one-to-one) is 0 and len(one-to-many) is 0
        return  removed_char_count
    else if len(one-to-one) is 0
        return HandleOneToMany(one-to-many)
    else
        remaining_GTs, remaining_Dets = CharacterRemovalProcess(one-to-one)
        return main(remaining_GTs, remaining_Dets)
    end
end
def InspectRelation(GTs, Dets):
    initialize one-to-one, one-to-many as array
    for GT in GTs
        Dets_intersect = filtered Dets intersecting GT
        if len(Dets_intersect) is 1
            one-to-one.append( [GT, Dets_intersect] )
        else if len(Dets_intersect) over 1:
            one-to-many.append( [GT, Dets_intersect] )
        end
    end
    return one-to-one, one-to-many
end
def HandleOneToMany(one-to-many):
    GTs, Dets = export GTs and Dets from one-to-many
    GT = closest one to left top of image among GTs
    Det = one with the highest area recall on GT among Dets
    remaining_GTs, remaining_Dets = CharacterRemovalProcess(GT, Det, GTs, Dets)
    return Main(remaining_GTs, remaining_Dets)
end
```

GT-to-detection. If there are multiple GTs corresponding to one detection, then the iteration order is predetermined.

$$k \in \{1, .., n\} \tag{1}$$

$$f(k) = \sqrt{{X_k}^2 + {Y_k}^2} \tag{2}$$

$$D = \underset{x}{\operatorname{argmin}} f(x) \tag{3}$$

For each iteration, the index of remaining GT polygons is k. The coordinates of k-th polygon centroid are X_k, Y_k. The distance between origin and the k-th polygon centroid is f. D indicates the nearest polygon to left-top of the image, that is subjected to next processes.

Algorithm 2: PopEval evaluation metric, part 2 of 2

```
def CharacterRemovalProcess(GT, Det, GTs, Dets):
    GT_text, Det_text = Extract texts from GT, Det
    for Det_character, Detchar_index in Det_text
        if  Det_character in GT_text
            GTchar_index = the leftmost character index of GT_text matching Det_character
            delete a character of GTchar_index from GT_text
            removed_char_count += 1
            if GT_text is empty
                delete GT object from GTs
            end
        end
        if Det_character is last of Det_text
            delete Det object from Dets
        end
    end
    GTs, Dets = the deleted GT and Det objects should be excluded from GTs, Dets
    return GTs, Dets
end
```

B. Handling the One-to-Many Relations

When the one-to-many relations are encountered in the course of GT-to-detection, it is required to pick one of the detected polygons matching the GT polygon in one-to-one relation. To solve this problem, PopEval adopted a recursive procedure which repeats inspecting every relation among GT polygons and detected polygons then preferentially taking out the obvious one-to-one relations, until only one-to-many relations remain. While the recursion is repeatedly executed, the previous one-to-many relations can be converted to one-to-one relationship in the current operation because obvious relations were wiped out in previous execution. For the last remaining one-to-many relations, the area recall of each detected polygon is adopted as the determinant to select one detected polygon that is matched with the GT: the larger the area recall, the higher the priority. Generally, only one detected polygon is subjected to character removal process. However, when there are multiple detected polygons with the same highest area recall, these are subjected to the process while being weighted as a reciprocal of the number of the subjected detections.

C. Character Removal Process

For character removal, the basic unit of comparison is a set of polygon area and the transcript. PopEval compares the units among GTs and predicted results. If the polygon areas of GT and predicted result overlap to each other, the transcriptions of GT and predicted result are compared, then the overlapped character is removed one by one in the predetermined order.

The principle of PopEval algorithm is to adopt the way a human evaluates. In the principle, there are two rules about the order in which characters are removed.

First, in the unit, the removal iteration of characters is conducted in the direction how characters are read. In this study, as the test dataset was in English, the iteration order of character removal was left to right. For (B) in Figure 1, the GT transcript is "POPEVAL", and the recognized transcripts are "OP", and "EVAL". According to *handling the one-to-many relations*(III-B), a unit of "EVAL" is subjected to the *character removal process*(III-C) first. The transcript of GT is removed from the characters of "EVAL" in order from left to right, then the transcript of GT becomes "POP".

Second, if a character of recognized transcript corresponds to multiple characters in the GT, the criteria to remove one character stays the same as above, following the direction in which characters are read in the language. Continuing from the above example where the transcript of GT became "POP", and last recognized transcript was "OP". According the character removal order, "O" is removed first, then "P" will drop out of transcripts. However, when "P" is removed, there are two candidates of character removal in GT transcript "POP", the first and third. As the direction determined above, the first character of the "POP" is picked to be matched and removed together with "P" of recognized transcript. As final result, the remaining GT transcript is "P" and there is no remains in recognized transcripts.

Recall and precision are calculated from the lengths of the initial GT and recognized transcript, and the number of removed characters. In this example, the length of the initial GT is seven, the initial total length of recognized transcript is six, and six characters were removed. The precision and recall are 1.0 and 0.8571, respectively.

IV. Experimental Result

A. Inspection on the Case of Concern in the PopEval

Currently, there is no perfect metric in the evaluation of OCR [9], and it is important that which metric is actually more accurate. Since PopEval is a method to remove overlapped character components between objects, there is a room for concern that it may not reflect the permutation problem in which the recognized transcript has different character arrangement compared to GT transcript. Therefore, the permutation problem was monitored by inspecting how frequently the problem occurs on state-of-the-art recognition models: attentional scene text recognizer with flexible rectification(ASTER) [10], and gated recurrent convolution neural network for OCR(GRCNN) [11]. Table 1 shows the occurrence of permutation problems on recognition models and test datasets. Test datasets of ICDAR2013 and ICDAR2015 were inspected. The permutation problem is defined as below:

Table I
Among the recognition results which composed of the same alphanumeric components as GT, the proportion that does not exactly match GT.

	ICDAR2013	ICDAR2015
ASTER	0.00%	0.05%
GRCNN	0.00%	0.14%

1) The transcripts of GT and recognition have the same character component.
2) The character arrangements of the transcripts are different to each other.

The survey showed that the permutation problem has rarely occurred. In the case review of the results, it is found that common permutation occurrences on the two models were both caused by a typing error in the GT. Subsequently, there was no permutation problem in test dataset of ICDAR2013 and out of 1811 images in total, the permutation occurred once with ASTER model, twice with GRCNN model for the dataset of ICDAR2015. Therefore, the occurrence of permutation problem is rare, considered as scarcely impinge on evaluation.

B. Occurrence of One-to-Many and Many-to-One Relations

IOU thresholding and exact text matching methods only accept one-to-one relations [9]. To inspect the errors caused by ignoring one-to-many and many-to-one relations, PixelLink [12] and EAST [13] were adopted as the detection model, and the ASTER and the GRCNN recognition models were trained as recognition model and made to predict on the test datasets of ICDAR2013 and ICDAR2015 competition.

One-to-many and many-to-one relations were counted on Table II under below criteria .

1) One-to-Many(split): If the recognized transcript of either GRCNN and ASTER is included in the GT transcript and the area precision of detection and GT boxes is greater than 0.5, the detection box is counted as a box in one-to-many relation.

2) Many-to-One(merge): If a GT transcript is a part of the recognized transcript of either GRCNN and ASTER and the area recall of detection and GT boxes is greater than 0.5, the GT box is counted as a box in many-to-one relation.

The assessment found that a non-negligible number of detection boxes and GT boxes were in one-to-many and many-to-one relations. Although there is ambiguity that the boxes in the relations match well each other in terms of shape and area, the transcript of the boxes is still valuable. In the approach ignoring these relations, all of the split detections and merged GTs are evaluated as false negatives. Additionally, since the relation assessment aforementioned relies on an imperfect recognition model, it is expected that

Table II
THE PROPORTIONS OF SPLIT DETECTIONS AND MERGED GTS AMONG THE TOTAL DETECTIONS AND GTS, RESPECTIVELY.

	Split Detections (one-to-many)	Merged GTs (many-to-one)
EAST - ICDAR2013	3.84%	1.46%
PIXEL - ICDAR2013	6.09%	3.29%
EAST - ICDAR2015	1.13%	1.54%
PIXEL - ICDAR2015	2.05%	0.35%

Table III
COMPARATIVE ANALYSIS OF POPEVAL F-SCORE FOR WORD LEVEL AND CHARACTER LEVEL ANNOTATION OF BENCHMARK DATASETS.

For ICDAR2013 Test Dataset

	Word Level	Character Level	Diff
EAST - ASTER	0.8649	0.8616	0.0033
PIXEL - GRCNN	0.8562	0.8531	0.0031
EAST - ASTER	0.8540	0.8513	0.0027
PIXEL - GRCNN	0.8552	0.8538	0.0014

For ICDAR2015 Test Dataset

	Word Level	Character Level	Diff
EAST - ASTER	0.8017	0.7991	0.0026
PIXEL - GRCNN	0.7696	0.7661	0.0035
EAST - ASTER	0.7792	0.7783	0.0009
PIXEL - GRCNN	0.8003	0.7986	0.0017

there will be more cases of one-to-many or many-to-one relations than the occurrences assessed.

C. PopEval's compatibility with benchmark datasets annotated at word-level and character-level

Since PopEval is an approach to evaluate the matched character component between GT and detection, it is most accurate when it is used on a character-level annotated benchmark dataset. Because benchmark datasets commonly have been annotated on word-level, we reannotated the test datasets of the ICDAR2013 and ICDAR2015 competitions on character level.

OCR models were evaluated at word-level and character-level then the compatibility between the results at character and word levels was investigated. Efficient and accurate scene text detector(EAST) and PixelLink were adopted as the detector model and ASTER and GRCNN were adopted as the recognition model. Therefore, the four detector-recognizer models were established, then evaluated on each of word-level and character-level benchmark datasets. As a F1 score, the harmonic means of recall and precision were calculated on Table III. The difference of F1 score between the evaluations on word-level and character-level datasets constantly stayed below 0.004. Considering the minor difference between the evaluations on word-level and character-level annotatated datasets, therefore, PopEval is compatible with the existing benchmark datasets which were annotated at word-level .

D. Correlations Between the End-to-End Evaluation Algorithms and Manual Qualitative Evaluation.

Since each existing evaluation algorithm has its own limitation, it is difficult to quantitatively determine which algorithm is more accurate. In this study, a qualitative evaluation was manually performed as a standard to compare the evaluation algorithms. For the qualitative evaluation, the participants used an assistant tool visualizing locations and transcriptions of GT and OCR result. Considering "do not care" marking of ICDAR2015 [2], the predicted polygons corresponding to the "do not care" markings were removed as preprocessing of the qualitative evaluation. The performance was evaluated as a character-oriented method by considering the errors of insertion, deletion, and substitution. A percentile of performance was marked as a five point scale: 0% to 20%, 1 point; 20% to 40%, 2 point; 40% to 60%, 3 point; 60% to 80%, 4 point; and 80% to 100%, 5 point;

To assess the correlation between evaluation algorithms and the manual qualitative evaluation, the average of three participants' scores and the results of following end-to-end evaluation algorithms were subjected to the assessment: the vocabulary-aided transcript matching with IOU over 0.5; the average precision with IOU over 0.5; the 1-NED; the PopEval using word-level dataset; and the PopEval using character-level dataset; For OCR model subjected to the assessment, because the Pixelink obtains an object by postprocessing, there is an ambiguity in calculating the confidence rate of the object for measuring average precision(AP). Therefore, the EAST as a detection model and both of the recognition models were subjected to the assessment, then there were two OCR models to be evaluated. For benchmark dataset, the test datasets of ICDAR2013 and ICDAR2015 were subjected to the assessment.

Pearson correlation was adopted to assess linear correlations between the manual qualitative evaluation and the end-to-end evaluation algorithms. As the result of the assessment, the PopEval was found to be the most similar to the manual qualitative evaluation in all cases. For ICDAR2013, the PopEval with character-level dataset showed very high correlation as 0.946 with the manual qualitative evaluation, nearly followed by the PopEval with word-level dataset. In Pearson correlation, coefficient above 0.8 means strong linear correlation in general. Although the correlation between PopEval and manual evaluation relatively decreased for EAST-GRCNN model with ICDAR2015, it still showed a strong correlation with the manual evaluation, followed by the other algorithms, and the traditional algorithms also showed lower correlation for EAST-GRCNN model than for the other.

This experiment showed that PopEval is the most correlated evaluation method with human qualitative evaluation among existing evaluation methods. Among traditional evaluation algorithms, the 1-NED showed the most correlation with manual qualitative evaluation.

Table IV
THE PEARSON CORRELATION COEFFICIENTS BETWEEN THE MANUAL QUALITATIVE EVALUATION AND THE END-TO-END EVALUATION ALGORITHMS FOR TWO OCR MODEL.

For ICDAR2013 Test Dataset

	Vocab	AP	1-NED	PopEval at word	PopEval at character
EAST - ASTER	0.7858	0.4595	0.8884	0.9305	0.9340
EAST - GRCNN	0.7910	0.4437	0.8800	0.9457	0.9461

For ICDAR2015 Test Dataset

	Vocab	AP	1-NED	PopEval at word	PopEval at character
EAST - ASTER	0.7776	0.5792	0.8124	0.9272	0.9213
EAST - GRCNN	0.6870	0.5410	0.7262	0.8221	0.8204

Vocab: vocabulary-aided transcript matching with IOU over 0.5; AP: average precision; 1-NED: normalized edit distance; PopEval at word: PopEval with word level dataset; PopEval at character: PopEval with character level dataset;

V. DISCUSSION AND CONCLUSION

As referred by Wolf and Jolion in [3], the drawback of object-oriented matching is the requirement that the bounding box wraps the actual text area tightly. For this reason, the rectangle approach scheme was only suitable for document images. In contrast to document scanning, however, extracting text from natural scene is much more difficult as the texts come with many varieties such as different orientations, varying aspect ratios or even skewed shapes. To account for these varieties, the four-vertices polygon approach was adopted as an annotation method recently. For recent interests of OCR, however, the four-vertices polygon method has the same limitation that the text area should be wrapped tightly, especially for curved texts. Therefore, a new annotation method with polygons of unlimited number of vertices is needed.

In benchmark dataset with polygon of unlimited vertices, the conventional approach is not appropriate, such as IOU, DetEval and average precision at a specific IOU. For GT of quadrilaterals, most of the relations between GT and detection were one-to-one, and the conventional criteria concepts considered only one-to-one relations ignoring the others. However, because the split and merge relation occurs more frequently with datasets with polygons of unlimited vertices such like Total-Text dataset [14], the concept of the object matching should be changed to reflect the actual performance.

In recognition, the vocabulary based evaluations are not adequate for wild scene text. Because the wild scene has varying texts such as unique nouns [9], dictionary based end-to-end evaluation is by its structure incapable of handling wild scene text. Even in strongly and weakly contextualised evaluations [2], the dictionary based evaluation has a limitation of not reflecting actual performance. Therefore, the current evaluations of recognition are based on edit distance and exact matching method.

When a recognition model recognizes a long string correctly, the model should be rated better than other models that recognized short strings. However, the exact matching method has its own drawback of not considering the various difficulty of each recognition because the method does not take into account partial correctness. The exact matching method causes underestimation of the model's actual performance, and the miscalculation depends on characteristics of benchmark dataset in use. Considering the above drawbacks, it is deemed desirable to approach the character-oriented evaluation rather than the object-oriented evaluation. Because the current OCR interests, such as multi-language transcripts, are more difficult to detect and recognize correctly than the previous tasks, the character-oriented evaluation is essential to evaluate the actual performance. In this aspect, the 1-NED was suggested as an end-to-end evaluation [8]. However, because it adopted IOU threshold as the criteria of object detection, this caused limitations due to threshold and ignoring split and merge relations.

Because the character-oriented evaluation requires character-level annotated dataset for accuracy evaluation, the character-level annotation should be provided as a test dataset in the future. Correspondingly, in order to develop PopEval, we newly annotated the existing benchmark datasets at character level. Although PopEval was devised to evaluate benchmark datasets annotated at character level, the evaluation method can be applicable to word-level benchmark dataset. The experimental results show that PopEval is compatible with word-level annotation, meaning PopEval can evaluate previous end-to-end tasks at character-oriented level without re-annotating the datasets at word-level.

The PopEval is a consistent performance evaluator for various benchmark datasets. In benchmark datasets annotated as unoriented rectangle box, the texts were not tightly wrapped by the ground truth annotations. This ambiguity necessitates conventional evaluation metrics to use variable thresholds for different benchmark datasets [3]. Different thresholds need to be applied to different benchmark datasets based on their characteristics, and incorrect results can be occurred in this process [9]. On the other hand, PopEval does not use the threshold method, but adopts pixel recalls between a GT and each detection, and this enhances the consistency of PopEval for various benchmark datasets.

The PopEval is the most human-like end-to-end evaluation method. Although the concept of the edit distance has been an effective method for recognition evaluation, in the aspect of end-to-end evaluation, the 1-NED contains the incomplete detection evaluation criteria caused by IOU concept.

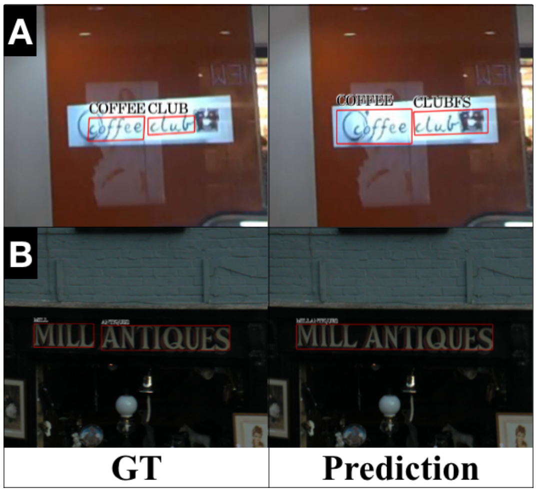

Figure 2. The representative evaluation cases showing limitation of traditional evaluation methods. The followings are evaluation results of each image; A: the IOU threshold can not catch the detected objects. 0.9090(PopEval), 0.0(1-NED), 0.0(AP); B: the merge relation occurred. 1.0(PopEval), 0.33(1-NED), 0.0(AP);

Through the correlation assessment between human qualitative evaluation and the algorithms, the PopEval showed much higher correlation with the human evaluation than the 1-NED. It means PopEval can handle the imperfection case of 1-NED, making its results more similar to those done with human evaluation.

The conventional evaluation methods such as DetEval, criteria of IOU and the edit distance have been adopted as the evaluation standard for a long time. Recently, it has been necessary to optimize the conceptual criteria of evaluation for new challenging tasks of OCR. In contrast to previous evaluation methods that are based on quadrilaterals, PopEval is able to handle polygons consisting of unlimited number of vertices. Above of all, the most innovative aspect is performing character-oriented evaluation with existing benchmark datasets annotated at word-level.

Further study and experiments are expected to enhance the integrity of PopEval. As with the imperfection of the existing evaluation methods, the permutation problem is a point of concern in PopEval. Although the experiment showed that the permutation problem rarely occurred, it is expected that concepts like the n-gram of BLEU can be applied to handle the sequence of characters in further study [15]. The character removal method, which provides compatibility with word-level datasets, is expected to contribute to more accurate model performance evaluation for future OCR tasks.

References

[1] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 Robust Reading Competition," in *2013 12th International Conference on Document Analysis and Recognition*, pp. 1484–1493, IEEE, 8 2013.

[2] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on Robust Reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, IEEE, 8 2015.

[3] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 8, pp. 280–296, 9 2006.

[4] K. Khurshid, C. Faure, and N. Vincent, "Word spotting in historical printed documents using shape and sequence comparisons," *Pattern Recognition*, vol. 45, no. 7, pp. 2598–2609, 2012.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, pp. 98–136, 1 2015.

[6] R. Gomez, B. Shi, L. Gomez, L. Numann, A. Veit, J. Matas, S. Belongie, and D. Karatzas, "ICDAR2017 Robust Reading Challenge on COCO-Text," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1435–1443, IEEE, 11 2017.

[7] A. Moffat and J. Zobel, "Rank-Biased Precision for Measurement of Retrieval Effectiveness," *ACM Trans. Inform. Syst*, vol. 27, no. 2, 2008.

[8] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17)," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1429–1434, IEEE, 11 2017.

[9] S. Long, X. He, and C. Yao, "Scene Text Detection and Recognition: The Deep Learning Era," 11 2018.

[10] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[11] J. Wang and X. Hu, "Gated Recurrent Convolution Neural Network for OCR," 2017.

[12] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting Scene Text via Instance Segmentation," 1 2018.

[13] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An Efficient and Accurate Scene Text Detector," tech. rep.

[14] C. K. Ch'ng and C. S. Chan, "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 935–942, IEEE, 11 2017.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, (Morristown, NJ, USA), p. 311, Association for Computational Linguistics, 2001.