An Impact of OCR Errors on Automated Classification of OCR Japanese Texts With Parts-of-Speech Analysis

Akihiro Kokawa*, Lazaro S.P. Busagala†, Wataru Ohyama*, Tetsushi Wakabayashi* and Fumitaka Kimura*
*Graduate School of Engineering
Mie University, 1577 Kurima-machiya-cho, Tsu, Mie, 514-8507
phone: +81-59-2319457 facsimile: +81-59-2319457
Email: see http://www.hi.info.mie-u.ac.jp/en/
†Sokoine National Agricultural Library/Computer Centre, Sokoine University of Agriculture, Morogoro, Tanzania
Email: busagala@suanet.ac.tz

***Abstract*—The technology of Optical Character Recognition (OCR) is used to generate texts in the process of digitizing print documents. Usually these texts need to be indexed and organized to simplify their access and retrieval. One of the powerful approaches in accomplishing this task is the use of Automated Text Classification. However, it is currently impossible for OCR technology to recognize all characters with an accuracy of 100%. We therefore propose the use of combined linguistic features in automated classification of OCR texts to formulate an informative feature set. The proposed method was experimentally evaluated using Japanese OCR texts. Empirical results indicate that the combination of linguistic features improved classification performance of OCR texts.**

***Keywords*-OCR Japanese text classification or categorization; Parts of speech analysis; Feature generation; Feature transformation; Combined Linguistic features.**

I. Introduction

The development of Information Technology has been increasingly changing the means of information exchange leading to the need of digitizing print documents [1]. There are generally two ways for digitizing print documents. These include (i) entering texts into a computer using a keyboard and (ii) using Optical Character Recognition (OCR) systems whereby text materials are extracted from digital text images.

The LDI project team [1] argues that the process of entering texts into a computer using a keyboard cost approximately 10-13 times more per page than using an uncorrected OCR at the Harvard University Library. Although automated classification of OCR Japanese texts is a feasible solution of organizing large volume of documents, such studies are rarely found in the literature.

The digitization process may involve creating digital images by a scanner and then generate ASCII texts by the use of OCR systems. However, it is impossible for OCR systems to give 100% accuracy. In other words OCR texts usually contain errors due to incorrect recognition of characters. These errors cause mis-representation of the documents for automated classification.

In information retrieval for example, OCR errors affect the system's reliability and efficiency. Many OCR error removal systems [2] focus on pages with combination of text and images. When dealing with only text applications, OCR error removal systems may delete important information of a particular text. Therefore there is a need to use OCR texts without applying any error removal process.

Traditionally, in languages other than Japanese, the bag-of-words (BOW) approach has been used in generating features to represent OCR texts [3], [4]. Usually BOW does not take into consideration of syntactic word categories. BOW includes all words even those which can cause over-fitting of document categories.

In the case of Japanese texts, the problem of classifying OCR generated texts can be tackled with a different view. Japanese being an agglutinative language poses a problem of not availing word delimiters. Unlike in a language such as English, words are not separated by white spaces. This has led researchers [5], [6] to invent methods that involve word segmentation for parts-of-speech (POS) estimation, which is also a difficult problem. Others [7] have proposed the use of specific Chinese characters (i.e. Kanji) in text classification. However, their work did not use linguistic features and standard corpora are missing to make a comparison to determine the best features. In [8] a POS filtering method in non-OCR texts is proposed where its performance did not outperform the use of all words with support vector machines (SVM).

Unlike conventional approaches, this work evaluates the use of linguistic features in improving the classification performance of OCR texts based on parts of speech analysis (POSA). In this research, the team used absolute frequency of syntactic word categories in this feature set, transformed them into a relative frequency and lastly, applied power transformation. This approach indicates that suitably selected POS features can improve classification effectiveness.

Section II introduces the proposed method for classification of OCR texts. In Section III experiments and learning algorithms are presented. Section IV provides the empirical

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.115

results while Section V provides a discussion on a brief survey of related works. Conclusions and present future directions are drawn in Section VI.

II. Linguistic Feature Combination

In this section we introduce the proposed method. We introduce features used in the experiments. The following subsections presents the details.

1) Feature Generation and Transformation: Let us consider a set of N sample texts, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ with n-dimensional text space. Assume that every textual document belongs to one of the C classes $\{\omega_1, \omega_2, ..., \omega_C\}$. Each text can be represented as a feature vector, $\mathbf{x}_l^{\omega_j} = [x_1\ x_2 \ldots x_n]^T$, whereby, n is the dimensionality (dimension of generated lexicon list), x_i is the absolute frequency(AF) of parts of speech (POS) which is the frequency value of i^{th} word or POS and T indicates transpose of a vector. In case of Kanji features x_i represents Kanji characters. Note that the superscript ω_j in the equations to follow is left out for the sake of simplicity. Absolute frequency tends to depend on text lengths leading into lower classification performance. This is due to the fact that text lengths differ within the same class of documents which make the learning process difficult.

In order to solve the problem of dependency on text length, the length variation in absolute frequency can be reduced by transforming AF to relative frequency RF as follows;

$$y_i = \frac{x_i}{\sum_{j=1}^{n} x_j} \tag{1}$$

In order to obtain normalized and Gaussian-like [9] distribution feature set transformation of the form:

$$z_i = y_i^v, \quad (0 < v < 1) \tag{2}$$

can be used. By substituting equation (1) into (2) we obtain a compact expression as

$$z_i = \Big(\frac{x_i}{\sum_{j=1}^{n} x_j}\Big)^v, \quad (0 < v < 1). \tag{3}$$

Features generated by equation (3) are called relative frequency with power transformation (RFPT).

2) Combination of POS Features: For simplicity, let us assume that there are two sets of feature vectors: (1) the feature set generated using nouns and (2) the feature set generated based on verbs present in a text. We denote nouns with a superscript u and verb features with a superscript v in equations (4) – (7). Consequently, we can define the feature vectors as

$$\mathbf{x}^{(u)} = \Big[x_1^{(u)}\ x_2^{(u)}\ \ldots\ x_{n_1}^{(u)}\Big]^T, \tag{4}$$

for the noun features. The verb features can be expressed as

$$\mathbf{x}^{(v)} = \Big[x_1^{(v)}\ x_2^{(v)}\ \cdots\ x_{n_2}^{(v)}\Big]^T. \tag{5}$$

The combination of these feature vectors can be defined as

$$\mathbf{Q} = \mathbf{x}^{(u)} \oplus \mathbf{x}^{(v)} \tag{6}$$

$$= \Big[x_1^{(u)} \ldots x_{n_1}^{(u)}, x_1^{(v)} \ldots x_{n_2)}^{(v)}\Big]^T. \tag{7}$$

Therefore, if we denote pronouns, adjectives and adverbs features with a superscript r, j and d respectively, then the combination of features can be expressed as;

$$\mathbf{Q} = \mathbf{x}^{(u)} \oplus \mathbf{x}^{(v)} \oplus \mathbf{x}^{(r)} \oplus \mathbf{x}^{(j)} \oplus \mathbf{x}^{(d)} \tag{8}$$

We use this technique to generate an informative features to get better classification performance. Equation (3) was applied to the composite feature set Q to form a normalized and Gaussian-like distribution.

3) Selection of suitable POS combination: We followed two steps in choosing suitable POS combination. Firstly we found out which POS contribute more in classification performance of OCR texts. In this step we removed one group of POS and then evaluate the classification performance to find out the hierarchy of the POS in describing a category's content.

Let us denote feature set of all POS as $\mathbf{A}$ and verb features as $\mathbf{x}^v$ the remaining feature vectors can be defined as

$$\mathbf{R} = \mathbf{A} \ominus \mathbf{x}^v. \tag{9}$$

Equation (9) can be used in removing other POS elements such as pronouns, verbs, adjectives and adverbs.

III. Experiments

A. Data for Experiments

In the experiments, we used articles published by Mainichi Newspaper of Japan from 1994 to 1999[1]. A total of ten categories i.e. Sports, International relations, Editorial, Economy, Home affairs, Culture, Reading, Science, Entertainment and Society. We drew a total of 1500 articles which were randomly split into 1000 and 500 articles for training and testing respectively. Each category was composed of 100 and 50 articles from training and testing accordingly. A word list from part-of-speech elements was generated from training data. Also Kanji character lists were generated for the feature vector formulations.

B. Pre-processing and Simulations

Test data documents from the Mainichi Newspaper collection were printed out in order to simulate the process of generating OCR texts. The print texts were digitized using a scanner into images of varying resolutions. These were converted into SHIFT_JIS texts by e-typist OCR software. These OCR texts were used as test data. All POS elements are extracted based on the morphological analysis using a Japanese text tool named *Chasen* [10]. Secondly, Kanji characters are extracted from Japanese documents.

[1]The website of the newspaper is http://mainichi.jp/

All Kanji found in the documents were also used to build a feature set and then these two sets were combined together to construct an informative feature set which improved the classification performance. In this case, experiments were done by excluding one type of the feature sets. The objective was to find out a subset of features that contributes much to distinctively represent Japanese documents.

The dimensionality was reduced by Principal Component Analysis (PCA) through which we retain a set of principal components with highest variance.

C. Learning Algorithms

Firstly, k nearest neighbor (kNN) is one of the best learning algorithm in terms of classification performance [11], [12]. The kNN algorithm relies on the concept that, given a test document $\mathbf{x}$, the system finds the k nearest neighbors among the training documents to estimate its *a posteriori* probability $P(\omega_j|\mathbf{x})$ for each category [13].

Secondly, Support Vector Machines (SVM) is the machine learning method which finds the optimal hyperplane with maximum margins. Interested readers can consult various work on SVM including references in [14].

We used the SVM^{Light} package [14] in the experiments. We divide each classification task into C binary classification problems and adopt the one against the rest strategy. [13], [14].

D. Performance Measures

In this section, we describe performance measures that were used in the evaluation process.

1) Word Recognition: To evaluate the accuracy of an OCR system we use the commonly used F-measure metric, which is equal to the harmonic mean of recall and precision. In this case, recall was calculated as a ratio of number of correctly recognized words to number of words in original text. Precision as a ratio of number of correctly recognized words to number of all words in OCR texts.

2) OCR Text Classification: We used the most used performance measure which is F_1 measure [15]. There are usually two averaging strategies. The first one is called micro-averaging. The second one is macro-averaging. The details for these strategies can be found in [16].

3) Significance of Improvement: McNermar's test is a statistical analysis technique that can validate the significance of the differences between two methods [17]–[20]. Let δ_1 and δ_2 be the first and second method, respectively. Testing the statistical significance can be done under a null hypothesis (H_0) that the two methods δ_1 and δ_2, would have the same error rate. The alternative hypothesis (H_A) is that method δ_2 has less errors than δ_1. If the test statistic z is greater than $z_{0.95} = 1.65$, a probability that there is a difference in observed performance of the two methods is less than the significance level $\alpha = 0.05$. In other words, if the statistical value $z > 1.65$, it would give a p-value that is less than the significance level $\alpha = 0.05$. This suggests that there is a statistical evidence that method δ_2 performs better than δ_1. Therefore we can reject H_0 in favor of H_A with reliability of 0.95 (95%). In this work we used one-sided hypothesis test as illustrated in Figure 1.

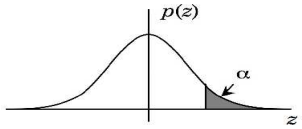


Figure 1. One-sided Hypothesis Test. $z_{0.95} = 1.645$ for $\alpha = 0.05$

IV. Results

Figure 2 is a graphical representation of Word Recognition Rate against classification performance for Mainichi dataset. It also shows that OCR text classification decrease with increase in OCR errors. The richer the feature set used to train a classifier the higher the classification performance. The combination of Kanji and POS features is more effective in OCR text classification than using the conventional features. As it can be observed in the Figure 2, the classification results without POS were lower. A comparison of empirical

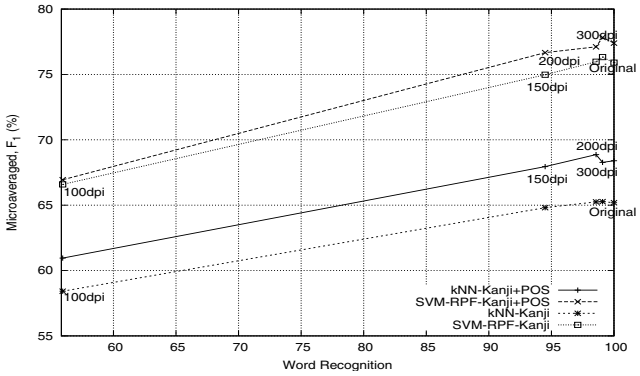
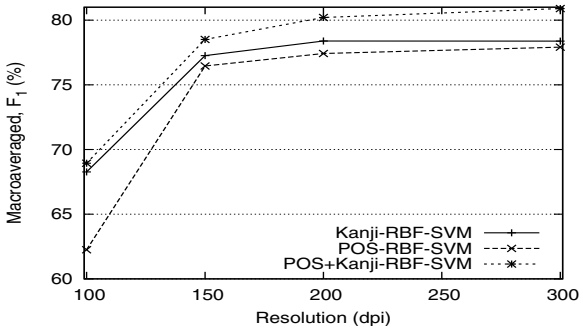
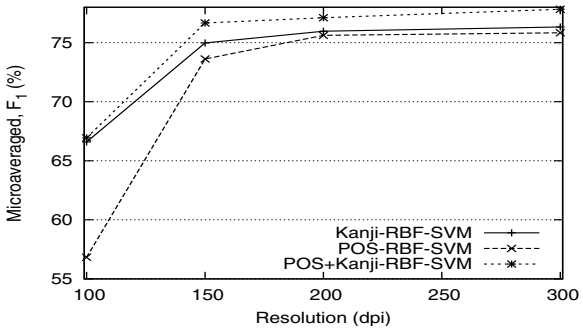


Figure 2. Effects of OCR errors in relation to Word Recognition (Micro-averaged F_1) in %) and Classification Performance (Micro-averaged F_1 in %) – Mainichi Dataset

results at various resolutions is presented in Figures 3. Detailed results are presented in Tables I and II. In Figure 3, it is observed that at lower resolutions of OCR images the classification of OCR texts is negatively affected. In this figure, it can be observed further that Kanji based features performed better than POS based features. The main reason for this might be the fact that large part of documents are composed primarily of Kanji based words. However the highest classification performance is observed when combined POS and Kanji features are used. This shows that the proposed method in this paper provides a more informative feature set than the conventional ones focusing on Kanji only. Figure 4 shows the result of excluding features that



(a) Macro-averaged Performance



(b) Micro-averaged Performance

Figure 3. Comparison of Results of RBF-SVM Method at Various Resolutions Using Mainichi Japanese Data

Table I
RESULTS OF RBF-SVM METHOD

DPI	Kanji		POS		Kanji+POS	
	Macro	Micro	Macro	Micro	Macro	Micro
100	68.27	66.59	62.25	56.83	68.94	66.94
150	77.26	74.98	76.46	73.62	78.51	76.67
200	78.39	75.97	77.42	75.62	80.21	77.11
300	78.38	76.33	77.91	75.84	80.90	77.83

belong to one type of parts of speech at a time. When general nouns were removed from the rest of the feature sets, the classification performance dropped down moderately. This means that general nouns, to some extend are important in representing Japanese texts distinctively. When adjectives were removed the performance of classifiers was generally high. In other words adjectives cause over-fitting between document categories because they can be found in most of the documents irrespective of the document categories. It can be said that adjectives do not contribute in distinctively representing documents in their respective categories. Furthermore we found that Kanji features are very important in distinctively representing Japanese texts.

In general, the figures show that Kanji feature sets are more robust to OCR errors than POS, and Kanji+POS features improve the robustness of the POS Features. The results for statistical test of significance of improvements are

Table II
RESULTS OF kNN METHOD

DPI	Kanji		POS		Kanji+POS	
	Macro	Micro	Macro	Micro	Macro	Micro
100	61.12	58.42	52.98	48.76	61.76	60.95
150	68.07	64.81	71.55	69.72	71.75	67.95
200	68.74	65.25	73.29	70.92	71.57	68.87
300	69.55	65.277	73.27	71.74	71.16	68.27

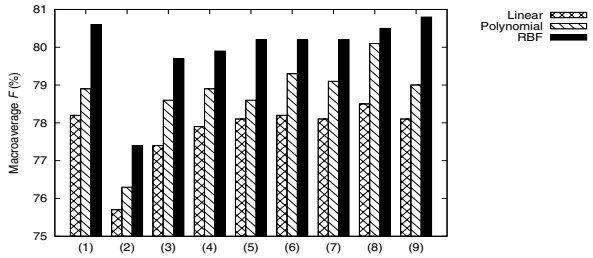


Figure 4. Performance resulting from excluding features of one type of parts of speech (POS) – Mainichi Dataset. Note: (1) No exclusion (Kanji+POS); (2) Kanji; (3) General Noun; (4) Sahen Connective Noun; (5) Proper Noun; (6) Adjective Verb; (7) Adverb; (8) Verb (9) Adjective

presented in Table III. It can be seen that the improvements were statistically significant with significance level $\alpha = 5\%$ for resolution 150dpi or less and with $\alpha = 10\%$ for 200dpi and 300dpi.

V. RELATED WORKS

The literature rarely shows research work done previously on parts-of-speech analysis in relation to Automatic Text Classification. To the best of our knowledge, this work of parts-of-speech analysis in relation to classification of Japanese OCR texts is appearing for the first time. Therefore the works we survey and present here, differ from this paper.

Murata et al. [4] proposed the use of transformed features for English OCR texts. Their work shows that using transformed features generated using the *bag-of-words* approach improved the classification performance. Our work in this paper differs from the work in [4] as we do not focus on *bag-of-words* as feature set. The work in [3] used *bag-of-words* which were untransformed features. They classified multi-page document by a hybrid naive Bayes HMM approach.

Watanabe et. el [7] have proposed the use of specific Chinese characters (i.e. Kanji). He extracted these Kanji by χ^2 method. The proposed method in this paper is different from the work in [7]. First of all we do not focus on specific Kanji. Instead we use all Kanji to represent Japanese documents. We then combine the extracted Kanji features with all POS. We make comparison and find that the proposed method is empirically better.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented a technique for linguistic features combination in Automated Classification of OCR

Table III
MCNERMAR'S TEST (KANJI V.S. KANJI+POS)

Resolution	100dpi	125dpi	150dpi	200dpi	300dpi
p-value(%)	0.0301	0.111	0.379	5.82	5.71

texts. It includes POS analysis. OCR errors negatively influence the learning algorithms. The higher the number of errors the lower the classification effectiveness due to inability of the learning algorithms. The focus on specific Kanji as reported in literature was outperformed by the proposed method.

The proposed approach in this paper has given promising results. Rather than using the conventional approach i.e, *bag-of-words* or specific Kanji, POSA and suitably combining other elements can improve classification performance. Kanji takes a great role in improving the performance of classifiers especially when combined with linguistic features. It is also noted that the Kanji feature set is more robust to OCR errors than POS, and Kanji+POS features improve the robustness of the POS features as confirmed by statistical test of significance.

Future work includes using experiments using more data samples. It is also interesting to assess the impact of OCR errors on POS tagging in future.

REFERENCES

[1] S. Chapman, "Measuring search retrieval accuracy of uncorrected OCR: Findings from the harvard-radcliffe online historical reference shelf digitization project." Harvard: Harvard University Library, 2001. [Online]. Available: http://preserve.harvard.edu/pubs/ocr_report.pdf

[2] K. Taghva, T. Nartker, A. Condit, and J. Borsack, "Automatic removal of "garbage strings" in OCR text: An implementation," in *5th World Multi-Conference on Systemics, Cybernetics and Informatics*, Orlando, Florida, 2001.

[3] P. Frasconi, G. Soda, and A. Vullo, "Text categorization for multi-page documents: A hybrid naive bayes hmm approach," in *In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. ACM Press, 2001, pp. 11–20.

[4] M. Murata, L. S. P. Busagala, W. Ohyama, T. Wakabayashi, and F. Kimura, "The impact of OCR accuracy and feature transformation on automatic text classification," in *Document Analysis Systems*, 2006, pp. 506–517.

[5] M. Nagata, "A stochastic japanese morphological analyzer using a forward-dp backward-a* n-best search algorithm," in *Proceedings of the 15th conference on Computational linguistics*, 1994, pp. 201–207.

[6] M. Nagata, "A part of speech estimation method for japanese unknown words using a statistical model of morphology and context," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 277–284.

[7] Y. Watanabe, M. Murata, M. Takeuchi, and M. Nagao, "Document classification using domain specific kanji characters extracted by χ^2 method," in *Proc. of 16th International Conference on Computational Linguistics IEICE The Institute of Electronics, Information and Communication Engineers*, 1996, pp. 794–799.

[8] H. Taira and M. Haruno, "Feature selection in SVM text categorization," in *AAAI '99/IAAI '99: Proceedings of the 16^{th} national conference on Artificial intelligence and the 11^{th} Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, 1999.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

[10] Y. Matsumoto, K. Takaoka, and M. Asahara, *ChaSen Morphological Analyzer version 2.4.0 User's Manual*, 2007.

[11] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the 12th International Conference on Machine Learning (ICML)*, Washington DC, 2003), pp. 616–623.

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[13] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the Twenty-First International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 42–49.

[14] T. Joachims, *Learning to classify text using support vector machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers Boston Dordrecht London, 2001.

[15] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol. 1, pp. 67–88, 1999.

[16] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[17] T. G. Dietterich, "Approximate statistical test for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[18] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 532– 535, glasgow.

[19] C.-M. Tan, Y.-F. Wang, and C.-D. Lee, "The use of bigrams to enhance text categorization," *Information Processing and Management: an International Journal*, vol. 38, no. 4, pp. 529–546, 2002.

[20] T.-Y. Wang and H.-M. Chiang, "Fuzzy support vector machine for multi-class text categorization," *Inf. Process. Manage.*, vol. 43, no. 4, pp. 914–929, 2007.