The Effect of OCR Errors on Stylistic Text Classification

Sterling Stuart Stein
Linguistic Cognition Lab
Computer Science Dept.
Illinois Institute of Technology
3300 South Federal Street
Chicago, IL 60616-3793
stein @ ir.iit.edu

Shlomo Argamon
Linguistic Cognition Lab
Computer Science Dept.
Illinois Institute of Technology
3300 South Federal Street
Chicago, IL 60616-3793
argamon @ iit.edu

Ophir Frieder
Information Retrieval Lab
Computer Science Dept.
Illinois Institute of Technology
3300 South Federal Street
Chicago, IL 60616-3793
ophir @ ir.iit.edu

ABSTRACT

Recently, interest is growing in *non-topical* text classification tasks such as genre classification, sentiment analysis, and authorship profiling. We study to what extent OCR errors affect stylistic text classification from scanned documents. We find that even a relatively high level of errors in the OCRed documents does not substantially affect stylistic classification accuracy.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.7.5 [**Document Capture**]: Optical Character Recognition

General Terms: Experimentation

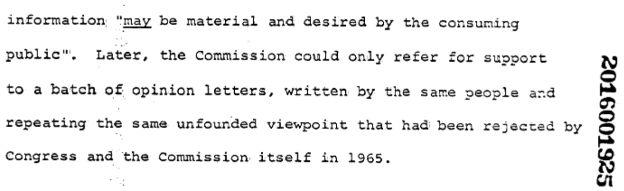
Keywords: OCR, OCR errors, text classification

1. INTRODUCTION

Recently, interest has grown in *non-topical* text classification tasks such as genre classification, sentiment analysis, and authorship profiling. Research on these problems, like work on 'classical' topic-based text analysis, has focused mainly on electronically produced digital documents. Real-world applications of automated stylistics in litigation, national security, and humanities scholarship require analysis of real paper documents which have been scanned and digitized via OCR. However, even the best OCR is not perfect, and introduces many transcription errors.

We present results of the first study we know of to evaluate the performance of style-based text classification on a corpus of OCR-processed texts (*OCR*), comparing classification accuracy to hand-corrected (*Correct*) versions of the same texts. We study text classification by genre (research reports, memos, etc) in tobacco industry documents. The parallel question has been previously investigated for the case of topic-based information retrieval; Taghva and Coombs [1] found that a search engine could be made to work well over OCR documents by accounting for the types of errors that it introduces. They ran misspelled words through an OCR-specific spell-checker and indexed the returned words based on a function of their probabilities.

Our evaluation is part of a larger project to develop a text collection [3] and integrated prototype for complex document information processing (CDIP), dealing with scanned documents that contain non-textual items as well as printed text. Our initial results show that OCR errors, though many, have little to no effect on the classification of the type of text.



information "may be material and desired by the consuming
public". Later, the Commission could only refer for support
to a batch of opinion letters, written by the same people and
repeating the same unfounded viewpoint that had been rejected by
Congress and the Commission itself in 1965.

2016001925

(a) Section of original document image

```
in.formation˜ ’˜.m_˜@_˜ be material and desired by the consuming
public’". Later, the Commission could only refer for support
to a batch o˜.opinion letters, written, by the same people and
repeating the. same unfounded viewpoint that had˜ been rejected by
Congress and the Commission. itself in 1965.
```

(b) OCR-extracted text

```
information "may be material and desired by the consuming
public". Later, the Commission could only refer for support
to a batch of opinion letters, written by the same people and
repeating the same unfounded viewpoint that had been rejected by
Congress and the Commission itself in 1965.
```

(c) Hand-corrected text

Figure 1: A document as an original image, after OCR, and after being corrected.

2. CORPUS

The corpus used in this study is composed of documents from the Legacy Tobacco Documents Library (http://legacy.library.ucsf.edu/) which we are using to build our IIT CDIP testbed. Each scanned document was run through OCR; there are 646 documents whose OCRed text was hand-corrected. Each document has a variety of metadata, including the type of the document, such as "Memo" or "Scientific Report"; it is these categories that we will attempt to predict in stylistic classification. The form of these documents can be seen in Figure 1.

In the raw data, there are many different such text-type labels. There is inconsistency in the labeling of each category, such as "Other Report" and "Report, Other". We combined these labels manually into 9 main text-types; documents of types that occurred fewer than 10 times in our corpus were removed. This left 326 total documents. The summary of the corpus can be seen in Figure 2. To measure the distance between *OCR* and *Correct*, we used Levenshtein distance [2]

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

Text type	**# Docs**	**Avg. Dist.**
Advertisement	13	0.60
Table, etc.	15	0.61
Correspondence	16	0.28
Published Doc	19	0.35
Press Release	28	0.14
Science	29	0.51
Other Report	42	0.31
Report	75	0.28
Memo	89	0.32
Total	326	0.33

Figure 2: Composition of the corpus. Distance between OCR and corrected versions was measured as the average edit distance between the texts as strings normalized by the document length, treating consecutive whitespace as a single space character.

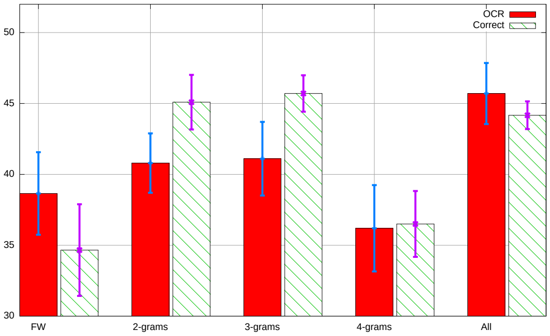


Figure 3: The 10-fold cross-validation accuracy with error bars for various feature sets. Note that considering error, there is no significant difference between *OCR* and *Correct*.

normalized by the length of the correct version. It should be noted that the OCR is reasonably accurate for text in paragraphs, however it is easily confused by headers. Headers were not removed.

3. METHODOLOGY

We applied a Support Vector Machine (SVM) learning method to build classification models. As input features, we used several types of numeric vectors, computed as the relative frequencies of textual attributes in each text. Probably the most common type of feature for stylistic text classification are function words, which were shown to be useful in many studies. Another type of feature that can be useful are character n-grams [4]. We compared results for both types of features separately.

For function words, we used a predefined list of English function words and computed the per-word frequency of each function word in each text as input features. For n-gram features, all n-grams (for $n \in \{2, 3, 4\}$) were counted, and the most common 1000 in the corpus overall were identified. Their counts were normalized for the length of each text and used as input features. For both *OCR* and *Correct*, these features were run through WEKA's SMO SVM [5] using the default settings with 10-fold cross-validation.

4. RESULTS

Overall results for different feature sets can be seen in Figure 3, all around 35-45%. Note that baseline accuracy for using the majority class would give 23%; so we are doing much better overall. The error bars shown are the standard error across cross-validation folds. More to the point, it is clear that text-type classification accuracy for *OCR* is not much lower than that for *Correct*, and is actually slightly higher for function words, but not significantly.

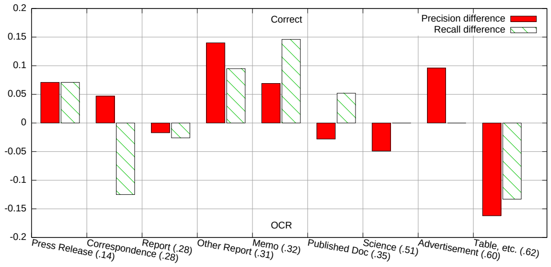


Figure 4: The difference in precision and recall of *OCR* and *Correct* for 2-gram features. Bars are grouped by text type and in order of increasing average distance. Positive means that *Correct* did better than *OCR*.

In figure 4, we show the differences in precision and recall between *OCR* and *Correct* for the various document types. No clear pattern emerges, but our corpus is too small to make any definitive statements.

5. DISCUSSION

We have found that stylistic classification accuracy is not significantly harmed, if at all, by OCR errors. In some of the cases, it even appeared slightly better. This illustrates how close the two data sets are, despite the errors.

Even though *OCR* contained many character-level errors when compared to *Correct*, the accuracy of the stylistic classification was comparable. This result argues that computational stylistics should be applicable to scanned document collections without much modification, although further work will be needed to examine different sorts of style-based text classification problems.

Acknowledgments

This work supported in part by an ARDA Challenge Grant.

6. REFERENCES

[1] J. C. Kazem Taghva. Hairetes: A search engine for ocr documents. *Intl. Workshop on Document Analysis Systems*, pages 412–422, August 2002.

[2] V. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.

[3] D. D. Lewis, S. Argamon, G. Agam, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *SIGIR-06*, 2006.

[4] B. K. O. Uzuner. A comparative study of language models for book and author recognition. *Springer-Verlag GmbH*, page 969, 2005.

[5] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham. Weka: Practical machine learning tools and techniques with java implementations, 1999.