EasyChair Preprint

№ 573

"How good is good enough?" Establishing quality thresholds for the automatic text analysis of retro-digitized comics

Rita Hartel and Alexander Dunst

October 11, 2018

"How good is good enough?" Establishing quality thresholds for the automatic text analysis of retro-digitized comics

Rita Hartel and Alexander Dunst

Paderborn University, Warburger Straße 100, 33098 Paderborn, Germany
rst@upb.de, dunst@mail.upb.de

Abstract. Stylometry in the form of simple statistical text analysis has proven to be a powerful tool for text classification, e.g. in the form of authorship attribution. When analyzing retro-digitized comics, manga and graphic novels, the researcher is confronted with the problem that automated text recognition (ATR) still leads to results that have comparatively high error rates, while the manual transcription of texts remains highly time-consuming. In this paper, we present an approach and measures that specify whether stylometry based on unsupervised ATR will produce reliable results for a given dataset of comics images.

Keywords: Graphic Novels, OCR, ATR, Automatic text analysis

1 Introduction

1.1 Motivation

Research on comics has undergone sustained growth over the last two decades in several disciplines and has now become a highly diverse field of inquiry. Although there are wordless and abstract comics, the medium's complex combination of words and images in telling stories has drawn the most sustained interest. Recent advances in image analysis and the explosive growth of the digital humanities (DH) mean that considerable efforts are underway to advance the computational analysis of comics. In previous work, we compared the automatic analysis of comics images with automated text analysis and were confronted with the problem that the quasi-handwritten fonts often used in graphic novels constitute a major challenge for state-of-the-art automatic text recognition (ATR) systems, although approaches for improving the performance of such systems for comics do exist [1].

This challenge led us to the question: "How good is good enough?" In other words: do we need a nearly perfect text recognition in order to perform text analysis, or are there certain tasks (e.g. analyses based on a term-document matrix) that can be performed on automatically recognized texts up to a given quality of the recognition.

1.2 Our Project

Our interdisciplinary project analyzes the different aspects of "hybrid narrative", in our case mainly graphic novels, comics narratives in book length that include fictional and non-fictional stories and are usually aimed at an adult audience. Fully automated analyses of such graphic novels are not yet feasible (beyond recognizing text there are even more difficult challenges, such as the recognition of narrative characters or the point-of-view of a panel). Therefore, our project semi-automatically annotates a corpus of currently around 220 graphic novels, memoirs, and non-fiction, which we call the Graphic Narrative Corpus (GNC) with the help of the M3-Editor developed as part of our project [2].

2 Automatic Text Analysis for Graphic Novels

Stylometry has proven to be a powerful tool for classifying documents, e.g. for authorship attribution. Even in the late nineteenth and early twentieth century, simple stylometric measures such as word length statistics [3] were used to determine the authorship of parts of the bible or of Shakespeare's plays. Later approaches were based on type-token ratio, i.e., the ratio of 'unique' words relative to text length or on number of *hapax legomena* (i.e., words occurring only once) [4].

Today, approaches to authorship attribution consider different methods. On one side of the spectrum, there are sophisticated methods based on machine learning. On the other hand, there is stylometry in form of simple statistical analysis of text. Machine learning has the disadvantage that it requires comparatively large training sets. Therefore, it might be more applicable to questions such as genre distinction, where the relation between the number of different genres and genre representatives is better than in the case of authorship attribution (more authors than genres, but far less novels per author than per genre). As a consequence, most authorship attribution is based (at least partially) on simple lexical features that are taken to be representative of the individual word usage of an author. These statistical analyses include traditional bag-of-words text representation that researchers use for topic-based text classification (also referred to as term-document matrix) [5]. Therefore, for our analysis, we decided to use 'traditional' stylometric features. Examples of such stylometric features are word-length frequency distribution, sentence length, word or character n-grams, PoS (part of speech) or function words. Specifically, the term-document matrix, i.e., the frequencies of the most common words of a corpus within a document, is used to compute the stylometric distance of several documents and is found to be among the best features for authorship attribution [6, 7].

For many lexical features, text is considered as a bag of words (i.e. an unordered collection containing duplicates) rather than a sequence. Other techniques, including n-grams, consider context [8] but frequently do not perform better than simple word-

based features [9]. Furthermore, in this paper we are not interested in semantical analysis, as is the case in part-of-speech-taggers, for instance, but consider words as syntactical units with certain features (e.g. their frequencies).

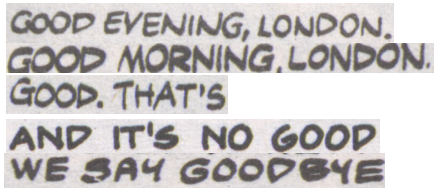


Fig. 1. Systematical error when recognizing the word GOOD in "V for Vendetta"

Looking at these features, we can see that errors made when automatically recognizing document texts might not constitute a serious problem. This is particularly true if these errors are made systematically: e.g., if a word w is always recognized as the wrong word v throughout the complete text (c.f. Fig. 1). (c.f. Fig. 1). These features may even benefit from systematical errors, if we consider that one author might use the same quasi-handwritten font throughout all of his work, whereas different authors will use different fonts (c.f. Fig. 2). That means that the wrong word v might occur only in the texts of author A and not in the texts of other authors.

Fig. 2. Different occurrences of word GOOD in our corpus

In this paper, we use a small sample of annotated pages from the GNC to determine if textual analysis based on the output of a given ATR system will produce reliable results.

2.1 Error rate measures

When evaluating the performance of systems for tasks like speech recognition, automatic translation, optical character recognition (OCR) or automated text recognition (ATR), two common measures are the character error rate (CER) and the word error rate (WER). For the two texts GT (or 'ground truth', the original text) and R (the recognized text), where R consists of n words, we can define the WER as the "normalized edit distance" of R to GT, i.e., the number of words of R that have to be substituted,

deleted or inserted in order to produce the original text GT, divided by the length n (in order to normalize the WER to be independent of the text length). Similarly, for the two texts GT and R, where R consists of m symbols, we define the CER as the number of symbols of R that have to be substituted, deleted or inserted in order to produce GT divided by the length m. CER is the more precise measure, i.e., typically, it holds that CER<WER for a document R. As discussed above, many stylometric features do not consider text as a sequence but as a bag of words. If most of the words are recognized correctly but their order was not assigned properly, this might lead to large CERs and WERs although the analysis is not affected, as they result e.g. in a very similar term-document matrix.

For our analysis, we propose a further error measure that we call the *bag error rate (BER)* that does not consider the order of words. Let GT and R be two texts, and let W be a set of words such that for each word gt ∈ GT ⇒ gt ∈W and for each word r ∈ R ⇒ r ∈ W. Furthermore, let $\text{freq}^D(w)$: W → ℕ be a function that assigns each word w of W the frequency of w within document D. Then the bag error rate (BER) is defined as:

$$BER := \frac{\sum_{w \in W} |freq^{GT}(w) - freq^{R}(w)|}{\sum_{w \in W} freq^{R}(w)}$$

In other words, for each word occurring in either GT or R, we compute the difference in the number of occurrences and calculate them for all words. Then, we normalize this sum by dividing it by the number of words of R. This calculation yields a measure that is robust against changing the order of words and reflects the idea of the term-document matrix. It also follows the idea of other features, for instance word-length distribution.

2.2 Quality measures for document distance

In order to decide if text recognition is of sufficient quality for an analysis that uses a term-document matrix, we used the following two evaluations: The analysis based on a term-document matrix considers the distance between documents in an n-dimensional space, where each dimension reflects the occurrences of a frequent word in the corpus. The smaller the distance between them, the more similar are the documents. Thus, a collection of documents should be considered of sufficient quality for analysis if each document is situated close to the corresponding document. The first evaluation – called PERC in this text – computes the distance between all documents that have undergone ATR to each other document. We then calculate what percentage of the other documents is closer to the corresponding original text than the recognized text. The smaller the percentile, the more suitable the recognized document can be considered for automated text analysis. This evaluation considers documents in isolation, that is, without considering its context in the form of all other pages of the same graphic novel.

The second evaluation – called COR in this text – considers the frequency vectors *fo* of the original document and *fr* of the recognized document. It then uses Spearman's Rank Correlation Coefficient to decide if the distances between the original document and all other documents can be correlated to the distances between the recognized document and all other documents. The Spearman correlation between two variables is

equal to the Pearson correlation between the rank values of those two variables. In other words, it compares the order of the variable values but ignores real values. The higher the coefficient (i.e., the nearer it is to 1), the more suitable the recognized document can be considered for automated text analysis.

3 Evaluation

The goal of our evaluation is to decide if the bag error rate (BER) is a good measure for selecting a graphic novel for automated text analysis. In order to calculate the BER, we need a ground truth in the form of the original text. Therefore, we also examine what percentage of a graphic novel needs to be manually annotated for the purpose of establishing a ground truth, in order to then compute the text's BER for further determination.

3.1 Method

For our evaluation, we used Tesseract 4 in LSTM mode without additional training to recognize the texts [10]. We ran Tesseract on a complete page of each GN. Note, that running Tesseract on complete pages results in much worse recognition rates compared to running Tesseract on speech bubbles only (we yielded a mean CER of 27%, a mean WER of 44% and a mean BER of 34% for speech bubbles, but only a mean CER of 69%, a mean WER of 82% and a mean BER of 43% for complete pages). We decided to consider complete pages only, as each identification of speech bubbles prior to further analysis would require (manual) detection of speech bubbles with a considerable effort and/or source of errors. In a second step, we compared each recognized page to the original text and computed the CER, WER and BER for each pair of pages. Furthermore, we ran a stylometric analysis with the help of the STYLO package within R [11]. The resulting term-document matrix was then analyzed with the help of our both evaluation methods (PERC – percentile of distance and COR –correlation of distances of corresponding documents). Finally, we checked if a correlation between the BER of a page and its PERC and COR value can be found, and what portion of a document has to be evaluated in order to yield a significant correlation between BER of that portion and PERC and COR for the complete GN.

3.2 Corpus

For our evaluation, we used the graphic narrative corpus (GNC) [2]], which was designed as a representative corpus as part of our project. As we need a ground truth in order to evaluate the results and the annotation of a graphic novel is very time-consuming (especially the transcription of the texts), only the part of our corpus that already has been completely annotated could be used. For our evaluation, we analyzed 13 graphic novels, written by different authors and belonging to a number of genres. In total, we analyzed 2,643 pages.

Graphic Novel	CER	WER	BER
A Contract With God	0.64	0.78	0.52
Batman – The Dark Knight Returns	0.74	0.88	0.53
Black Hole	0.55	0.71	0.27
City Of Glass	0.69	0.85	0.4
Fun Home	0.48	0.59	0.22
Gemma Bovery	0.94	0.97	0.73
Harvey Pekars Cleveland	0.96	0.99	0.74
Jimmy Corrigan	0.66	0.78	0.33
Our Cancer Year	0.76	0.89	0.43
The Complete Maus	0.65	0.88	0.48
The Diary of a Teenage Girl	0.97	0.99	0.91
V for Vendetta	0.68	0.86	0.46
Watchmen	0.63	0.81	0.29

Table 1. Graphic Novels used in our evaluation and there mean error rates

3.3 Results

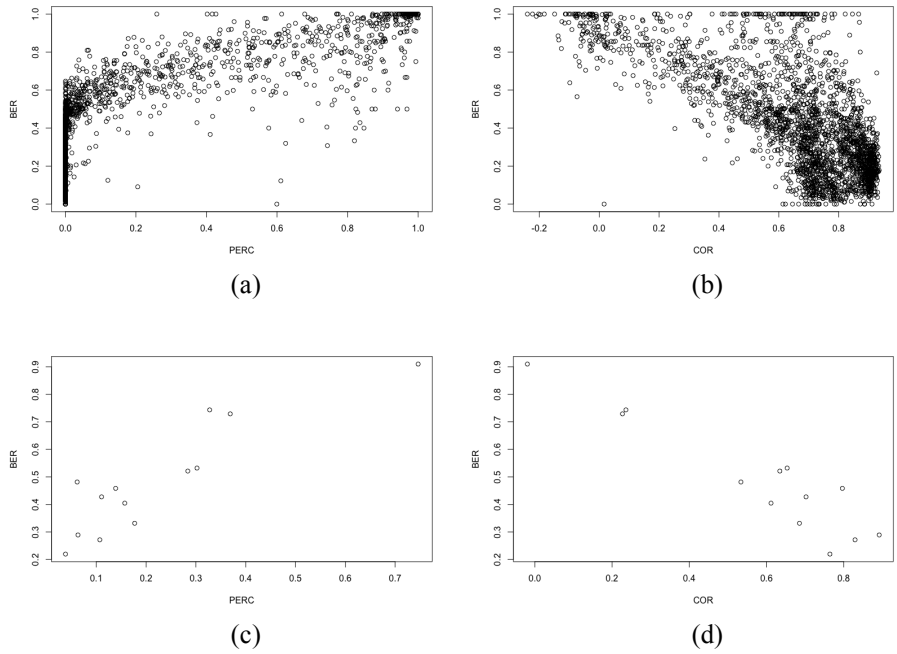
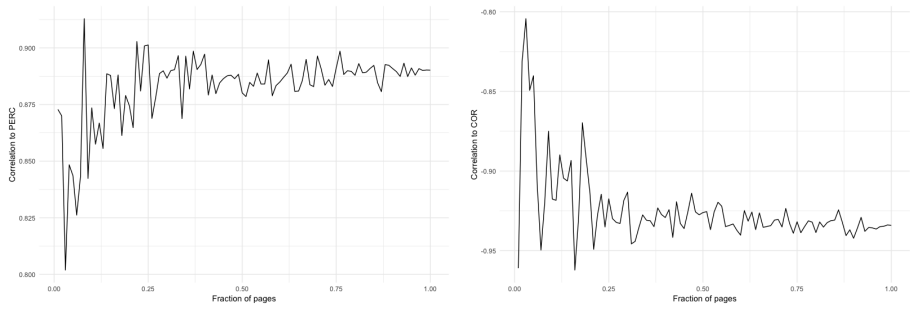


Fig. 3. Correlations for single pages between BER and PERC (a) or COR (b) and correlation s for mean values per graphic novel between BER and PERC (c) or COR (d)

Our evaluation shows that there is a strong correlation between the bag error rate and our two measures PERC and COR. As shown in Fig. 3, there is also a strong correlation between BER and PERC with a Pearsson's rank coefficient of 0.81 and a p-value of less than $2.2*10^{-16}$, as well as a medium strong correlation between BER and COR with a Pearsson's rank coefficient of -0.71 (i.e., the smaller the error, the better the Spearman's correlation of the document's distances) and a p-value of less than $2.2*10^{-16}$. If we aggregate the BER, PERC and COR for complete graphic novels, we reach even stronger correlations, with a rank coefficient of 0.89 (and $p<0.000046$) for BER/PERC and a rank coefficient of -0.93 (and $p<0.0000029$) for BER/COR. These results allow us to state that for our evaluation corpus, the BER of a complete graphic novel functions as a good estimator of the value of ATR for all stylometric analyses that are based on bag of words.

A BER of 0.4-0.5 seems to be a good threshold to yield documents, or graphic novels, with a Spearman's Rank Coefficient of more than 0.6 (or more than 0.8 even in many cases). Therefore, in our successive evaluations, we use the threshold of $BER<0.5$ to choose documents for automated text analysis.

As still we need a ground truth, in our second evaluation we compared the fraction of the graphic novel for which we computed the BER with the correlation coefficient of BER to PERC and BER to COR.



(a) (b)

Fig. 4. Progress of Correlation Coefficient for BER/PERC (a) and BER/COR (b) for growing fraction of pages for each graphic novel used for calculating the BER

Fig. 4 shows the results of this second evaluation. Even for small fractions of a graphic novel we can already establish strong correlations between the BER for this fraction and the PERC and COR for the whole GN. When choosing a random sample of around 10% of the pages, we can use the BER as a good estimator. When choosing around 25%, the correlation coefficient remains more or less stable.

We conclude our evaluation with a comparison of automated text analyses and a text analysis on the transcribed texts. We used the term-document matrix and performed a dimension reduction on it with the help of PCA in order to visualize the results. Fig. 5 (a) and (b) show the visualization of the term-document matrix for Charles Burns' *Black Hole* and Paul Auster, Paul Karasik and David Mazzuchelli's *City of Glass*. As

they were written by different authors and belong to different genres (coming of age and crime, respectively), these texts can be expected to possess distinct stylistic qualities. Part (a) shows the visualization of the automated text analysis, whereas part (b) shows the visualization of the analysis of the manually transcribed texts. As we can see in these figures, the two graphic novels can be distinguished quite well and the documents overlap only in a small part at the center of the plot.

Fig. 5 (c) and (d) show the visualization of the term-document matrix of two other graphic novels, *V for Vendetta* and *Watchmen*. Both were written by Alan Moore and, as a consequence, can be expected to possess similar text properties. Part (c) shows the visualization of the automated text analysis, whereas part (d) shows the visualization of the manually transcribed texts. In contrast to the graphic novels that we expected to be distinct stylistically, these two figures are quite similar: in both the documents of the two graphic novels overlap. Regions where only one of the two novels can be found are relatively minor in comparison. These examples support the results of our earlier evaluation: when choosing graphic novels that show a BER of less than 0.5, automated text analysis on the results of an automated text recognition system yields similar results as the analysis of texts transcribed manually in a highly time-consuming annotation process.

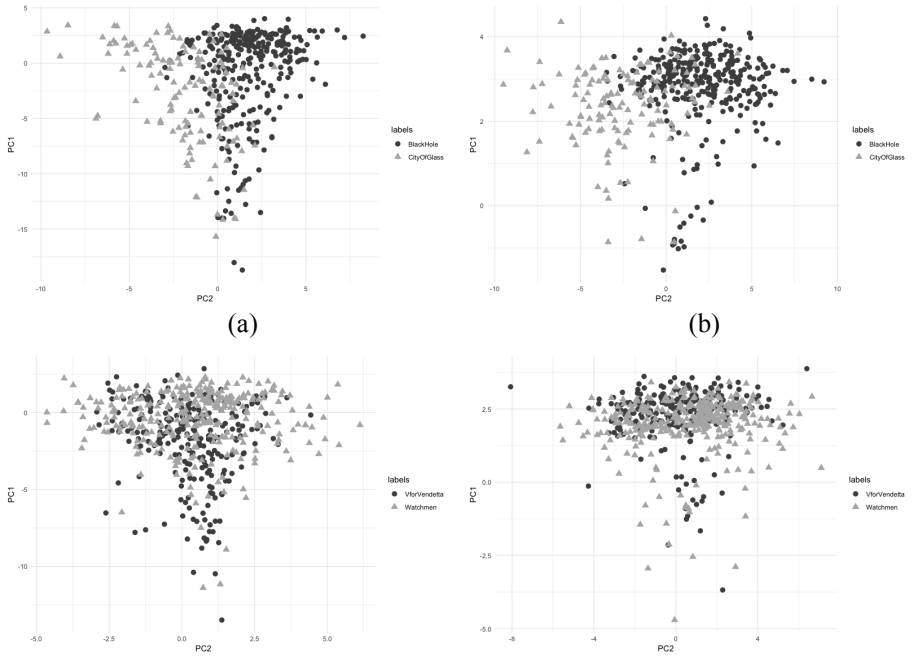


Fig. 5. Visualization of results for graphic novels with diverse texts (recognized (a) and original (b)) and for graphic novels with similar texts (recognized (c) and original (d))

4 Conclusion

In this paper, we presented an evaluation on the feasibility of automated text analysis based on the retro-digitized images of graphic novels, or book-length comics narratives. Within our evaluation, we could show that with the help of the bag error rate (BER) defined in this paper, we were able to establish a good estimator for the reliable stylistic analysis of graphic novels based on automatically recognized texts. In future work, it will prove an interesting task to extend the measures used for the stylometric analyses to other measures (e.g. n-grams and word-length frequencies). Currently, we are in the process of annotating around 10% of the pages for the entire GNC. Soon enough, we will thus be able to extend this research to automatically analyze large parts of the GNC and examine how well stylometric analysis can be used not only for authorship attribution but also for classification tasks, including genre distinction.

References

[1] C. Rigaud, J.-C. Burie und J.-M. Ogier, „Segmentation-Free Speech Text Recognition for Comic Books," in 2nd International Workshop on coMics Analysis, Processing, and Understanding, 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 2017.

[2] A. Dunst, R. Hartel und J. Laubrock, „The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities," in 2nd International Workshop on coMics Analysis, Processing, and Understanding, 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 2017.

[3] T. Mendenhall, „The characteristic curves of composition," Science, pp. 237-249, IX 1887.

[4] O. Y. de Vel, A. Anderson, M. Corney und G. M. Mohay, „Mining Email Content for Author Identification Forensics," SIGMOD Records, Bd. 30, Nr. 4, pp. 55-64, 2001.

[5] F. Sebastiani, „Machine learning in automated text categorization," ACM Computing Surveys, Bd. 34, Nr. 1, pp. 1-47, 2002.

[6] J. Burrows, „Word patterns and story shapers: The statistical analysis of narrative style," Literary and Linguistic Computing, Bd. 2, pp. 61-70, 1987.

[7] S. Argamon und S. Levitan, „Measuring the usefulnuss of function words for authorship attribution," in Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2005.

[8] F. Peng, D. Schuurmans und S. Wang, „Augmenting Naive Bayes Classifiers with Statistical Language Models," Information Retrieval Journal, Bd. 7, Nr. 3-4, pp. 317-345, 2004.

[9] C. Sanderson und S. Günther, „Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation," in EMNLP 2007,

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006.

[10] R. Smith, „An Overview of the Tesseract OCR Engine," in 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Paraná, Brazil, 2007.

[11] M. Eder, M. Kestemont und J. Rybicki, „Stylometry with R: a suite of tools," in Digital Humanities 2013, DH 2013, Lincoln, NE, USA, 2013.