A Rule-based Statistical Classifier for Determining a Base Text and Ranking Witnesses In Textual Documents Collation Process

Mohamadou Nassourou
Department of Computer Philology & Modern German Literature
University of Würzburg Am Hubland D - 97074 Würzburg
mohamadou.nassourou@uni-wuerzburg.de

Abstract

Given a collection of diverging documents about some lost original text, any person interested in the text would try reconstructing it from the diverging documents. Whether it is eclecticism, stemmatics, or copy-text, one is expected to explicitly or indirectly select one of the documents as a starting point or as a base text, which could be emended through comparison with remaining documents, so that a text that could be designated as the original document is generated. Unfortunately the process of giving priority to one of the documents also known as witnesses is a subjective approach. In fact even Cladistics, which could be considered as a computer-based approach of implementing stemmatics, does not present or recommend users to select a certain witness as a starting point for the process of reconstructing the original document.

In this study, a computational method using a rule-based Bayesian classifier is used, to assist text scholars in their attempts of reconstructing a non-existing document from some available witnesses. The method developed in this study consists of selecting a base text successively and collating it with remaining documents. Each completed collation cycle stores the selected base text and its closest witness, along with a weighted score of their similarities and differences. At the end of the collation process, a witness selected more often by majority of base texts is considered as the probable base text of the collection.
Witnesses' scores are weighted using a weighting system, based on effects of types of textual modifications on the process of reconstructing original documents.
Users have the possibility to select between baseless and base text collation. If a base text is selected, the task is reduced to ranking the witnesses with respect to the base text, otherwise a base text as well as ranking of the witnesses with respect to the base text are computed and displayed on a histogram.

Keywords: Textual document collation, Base text, Gothenburg model, Bayesian classifier, Textual alterations weighting system

1. Introduction

One of the main purposes of textual documents collation systems is to identify similarities and differences that exist between the documents under collation.
There are base text based collation and baseless collation systems. As explained in [1] baseless collation has got some disadvantages as far as visualization and explicit identification of types of textual changes are concerned.
The collation system used in this study is amply described in [1], which consists of tokenization, alignment, interpretation, and visualization units. It is during the interpretation phase that identification of types of textual alterations and their frequencies is carried out.
Some weights are assigned to each type of textual alteration based on the degree of difficulty to reconstruct the supposed original text from available witnesses.
Combination of weights and computed frequencies of types of textual modifications produces some scores for the collated documents, which will be used for ranking them with respect to a given selected base text.
A selected base text that share maximum similarities and minimum differences with all other witnesses is considered as the most probable base text that could be emended to reconstruct the original text.

This paper shows that eclecticism, stemmatics, and copy-text techniques could be combined, in order to identify a base text and grade remaining witnesses. First eclecticism is used to identify types of textual alterations, then stemmatics is employed to cluster the witnesses based on the result of eclecticism, and finally a copy-text is derived from the latter step.

In this study, an intelligent textual documents collation system implementing the extended Gothenburg model described in [1] for identifying a base text and ranking witnesses has been developed. The ranking is performed according to types of textual alterations that took place in the documents. Using a weighting system, a rule-based Bayesian classifier computes similarities and differences between a selected base text and remaining witnesses, and then it assigns some weighted scores to witnesses. Computed scores are used to assist identification of a base text, as well as witnesses closer to the base text.
Weights are assigned to each type of textual alteration according to the degree of complexity of reconstructing the original text from available witnesses. The system offers users the possibility to interactively modify the weighting system, and collate the documents again.

2. Motivation

Reasonably and consistently reconstructing a non-existing document from some available textual variants is an interesting scholarship activity. How to identify a variant that could serve as a starting point or a base text is not a trivial task. The identification of a base text is a kind of decision making process that requires support from the tools that one interacts with. For instance from religious texts perspective, one would like to know which one of the

variants could be considered as a base text, and witnesses closer to the base text, and so on.

Is it possible to establish a kind of ranking among the types of textual alterations? For example which one causes more difficulty in the process of reconstructing the original text?

The aim of the ranking is to effectively assist texts scholars while making their decision about textual variants.

3. Background

3.1 Textual criticism

Textual criticism is a process whereby a text approximating an unavailable original document is reconstructed. Usually the generated text is called a critical edition.
While some critical edition projects attempt reconstructing some lost texts by principally considering meaning and purposes of the texts, others such as religious texts reconstruction projects try bringing back the original texts word by word.
There are several traditional techniques used to reconstruct distorted or scattered textual documents from available witnesses, among them are eclecticism, stemmatics, and copy-text editing. Advantages and inconveniences of each technique are explained in [9].

With the fast growing computational capacity of computers, techniques from the discipline of biology namely cladistics, along with machine learning algorithms are also being used to implement the traditional techniques such as stemmatics.
By explicitly or indirectly selecting a base text each of these methods attempts producing a document, that might be thought to represent the unavailable original document.

From comparison of similarities and differences between available witnesses or readings, to generation of complete family trees, textual criticism manages to compose texts that make people believe their absolute resemblance to the original document.

During the old days these activities were carried out manually. But nowadays computer-assisted textual documents comparison is becoming a vulgar activity, because of the numerous existing software collation programs. Some software are for baseless collation, and others for base text based collation. With some adjustments some baseless collation software could be used for base text based collation as well.
For instance eclecticism and stemmatics are more suitable for baseless collation process, while copy-text is basically a base text based collation.
Whether a baseless or base collation is considered, one is compelled to analyze the result of the collation and directly or indirectly give priority to one of the witnesses or group of witnesses. The process of deciding which one of the witnesses is the closest to the original text remains a subjective issue.
Cladistics which could be viewed as computer-based stemmatics, does not select a witness that might be closer to the original text.

3.2 Textual Documents Collation Process

i. Collation Methodology

Textual documents collation refers to the process of measuring proximities between documents under collation. Proximity consists of similarities and differences that might exist among the collated documents. Therefore a collation system is the one that is able to locate common and unshared characters between textual documents along with their statistics. Ideally a textual documents collation system should not only look at proximities, but also at actual position of each character in the text. The position of characters is needed for the identification of types of textual changes that might have occurred in the documents.

ii. Definition of types of textual alterations

Referred types of changes in this study are changes that are not necessarily depending on human interpretation such as clarification, overwriting, fixation, and so on. For more information on textual document alterations see [2]. To the contrary these are changes which are identifiable through simple inspection and visualization of text documents. In fact these changes are susceptible of influencing the flow of the text. Deletions, additions, transpositions, and mutations are some of the textual alterations considered in this paper. Detailed information about these changes is provided in [1,2].
However it might be important to differentiate between substitution and mutation. A substitution does not in itself suggest how replacement of a text by another took place, while a mutation does. In fact a mutation is a term well known in medicine as well as in genetic algorithm. It suggests that something has been transformed fully or partially into something else. In [2] the case of mutations has not been mentioned, and I suppose substitution is meant for it.

iii. Extended version of Gothenburg model for collation process

In [1] an extended version of the Gothenburg model for textual documents collation process was designed and implemented. The model consists of tokenization, alignment, interpretation, and visualization units.
Figure 1 shows a diagram of the model.

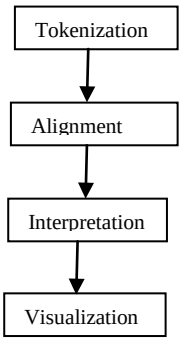

Fig.1 Document collation process steps

Explanation and reasons behind the design of fig 1could be found in [1].

3.3 Classifiers

There are several types of classifiers with each one having its own method of categorizing documents or objects. Among them we find statistical, functional, neural, decision trees, and fuzzy classifiers. The most widely used classifiers

are the statistical ones such as the Bayesian and distance-based classifiers. However some functional classifiers such as K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) are also intensively researched.
For this study, a Bayesian classification is selected because of its simplicity and effectiveness. More information about the selection of a Bayesian classifier for this study is found in [12].

Naive Bayesian Classifier

Bayesian classifiers assign a document X to a class C_i by applying the following Bayes' theorem:
$P(C_i|X) = P(X|C_i)\ P(C_i) / P(X)$

In traditional text classification, a document X is represented by an n-dimensional vector X = {x1, x2, . . . , xn}.
If the vector X is huge, processing time would be long to compute $P(X|C_i)$. Hence the naïve Bayesian classifier assumes that classes are conditionally independent.
In other words for a given vector X its entries x1, x2,…xn are conditionally independent of one another.
Mathematically this implies that:

$$P(\mathbf{X}|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i).$$

Given a document X, the classifier will predict that X belongs to the class C_i if and only if

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \qquad \text{for } 1 \leq j \leq m,\ j \neq i.$$

$P(C_i|X) = P(X|C_i)\ P(C_i) / P(X)$

For the current research P(X) is same for all classes, therefore only $P(X|C_i)P(C_i)$ needs to be maximized.
Moreover the class a priori probabilities $P(C_i)$ are not known, therefore I supposed they are all equal,
$P(C_1) = P(C_2) = \ldots = P(C_k)$.
Hence only P(X|Ci) needs to be maximized.

$P(X|C_i)$ is computed as follows:

$P(X|C_i) = w_f / d_f$

d_f is 1 since there is one document for each category.
w_f is the frequency of each type of textual alteration.

The frequency has to be weighted by multiplying it with its respective weight.
$P(X|C_i) = w_f$ x weight , where weight is defined in table 1.

So the probability of each witness denoted SCORE is the sum of all the $P(X|C_i)$ of the witness. Mathematically it could be written as follows:

SCORE(Witness) = $\sum w_f$ x weight

The Score will be used for selecting a base text as well as grading the witnesses.

4. System Architecture

The system uses CollateX library [10] for the generation of aligned tokens of the documents under collation.
Due to the fact that CollateX cannot accept XML as input, and for the time being there are no explicit methods for retrieving and distinguishing types of textual alterations, pre and post-processing steps were obviously unavoidable.

Fig. 2 shows the developed architecture, and its full description is provided in [1]. Shortly, it accepts some witnesses as input, preprocesses them by formatting and normalization, then collects results in form of a matrix, and then the post-processing unit is called for generating appropriate output format for visualization, as well as for further computation needed by the statistical classifier.
The statistical classifier component is responsible for the computation of frequencies of types of textual alterations, scores for each witness, and identification of a base text along with ranked remaining witnesses.

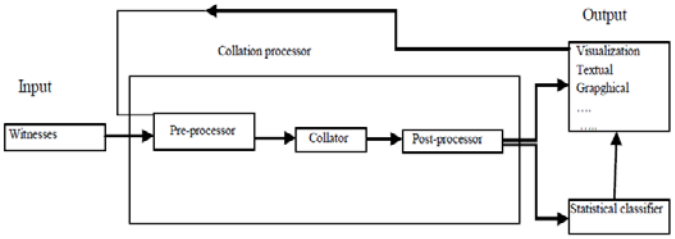

Fig.2. Architecture of an interactive textual documents collation system

5. The Approach

The determination of a base text and closer witnesses could be viewed as a categorization problem. In this study a Bayesian classifier is used to identify a base text, and rank remaining witnesses with respect to the base text. The approach to solve this problem is divided into two steps:

a. Designing a weighting system for grading types of textual alterations based on the degree of difficulties encountered, when reconstructing an original document from available readings or witnesses.
b. Use a Bayesian classifier to compute frequencies of types of alterations, and normalize them with the weighting system, in order to assign a score to each witness. Based on the scores select a base text, and rank remaining witnesses.

i. Weights Assignment

In this study only the four types of textual changes mentioned in section 3.2 have been considered, because these changes can be easily and adequately identified. Moreover they are the worst textual alterations that a text could undergo, since they affect the text flow, and end up giving a completely different interpretation of the text.

It might be important to mention that, the ranking of witnesses with respect to a possible base text is made according to human judgment rather than computer-based detection approach. In fact detecting types of textual alterations in some cases might be more difficult for a computer program than if it were carried out manually. For instance transpositions are computationally harder to uniquely identify.

Additionally the weighting system is also based on my personal experience with religious texts mostly Islamic texts, and some reasonable imagination of how types of

textual alterations could affect the reconstruction of a lost original document from available readings.

So, following is a table that summarizes the weights attribution to each type of textual modification.

Type of modification	Weight	Meaning
Mutation	4.5	highest mark i.e worst alteration
Deletion	3.5	very bad but less critical than mutation
Addition	1.5	bad, sometimes exaggerates the purpose of a text
Transposition	1.0	Misleading but could be easily detected and corrected

Table 1. weights assignment

Description of the table

Mutations are usually the worst case of textual document alterations. In fact it is very difficult to identify mutations if the base text is a priori unknown, because most of the time replacing words fit into the text flow, in such a way that, only extremely careful imagination could lead to their detection. Moreover it seems to be subjective to imagine something unknown that should fit into a text flow.

Deletions are the second worst case of types of changes. It is very confusing and sometimes polarizing, and even biasing to imagine missing words in a document. Just as mutations, they seem also to be very subjective.

Additions usually augment the meaning and aim of a text. It is therefore misleading, but might not necessarily contradict or alter the overall purpose of the text. Their detection might also be very tricky.

Transpositions are the easiest type of alterations to detect if a base text is known. If a base text is not known then only careful and reflective reading of the text could lead to their identification. Transpositions are considered less dangerous than other types of alterations, because individual words of the document and their number are intact, and their manual detection is not so critical.

Normalizing frequencies of types of textual modification is necessary, because the reconstruction of the original text depends mostly on those textual alterations that have been introduced in the witnesses.

Remarks:
It might be important to mention that, the chosen weight number for each type of alteration was arbitrarily made.
Therefore it might be necessary to carry out a formal and quantitative empirical study on effects of types of textual alterations on the process of reconstructing lost original texts, using univariate analysis method. This study could involve consulting and analyzing several critical edition projects with focus on types of textual alterations, difficulties, and complexities encountered in the process of eliminating them. It is hoped that quantitative estimation of those textual modifications could help design a more robust weighting system. This would be one of the themes for the future work.

ii. Classification Algorithm

The classification algorithm consists of two steps:

a. Identification of a base text
b. Detection of a closest witness, and ranking the witnesses with respect to the base text

a. Identification of a base text

The identification of a base text is an iterative process of the selection of a closest witness described in **b**.
Shortly it could be explained as follows:

1. Select one of the witnesses as base text, collate with remaining witnesses.
2. Find closest witness to the base text as described in **b**, and save the base text with the witness in a vector.
3. Repeat the process for every witness. At the end of the process one will have a vector containing pairs of base texts with witnesses.

The most frequent witness in the vector should be considered as a base text for all the witnesses. If the witnesses are equally represented, then one of them is selected randomly.

b. Selection of closest witness and ranking

The selection and ranking process is described with the following steps:

1. Set a base text among the witnesses if not identified.
2. Collate each witness with the base text by tokenizing and aligning them.
3. During the interpretation phase:
   a) compute frequency of each type of alteration
   b) normalize each frequency by multiplying with respective weight
   c) sum all normalized frequencies for each witness and save as its score
4. Sort the scores in ascending order
5. The witness closest to the base text is the witness with the smallest score. If there is more than one witness, then they are all considered as closest witnesses to the base text.

The closest witness could be considered as the most probable exact plagiat of the base text.

6. Results And Discussion

The classification is based on a rule-based weighted frequency system. Types of textual alterations found in each documents are computed and normalized using a weighting system based on some rules set according to human judgment of degree of complexity to reconstruct lost original texts.
Results obtained show that the weighting system works as well as the rules.
The accuracy of the classifier has been measured using precision and recall metrics.

Some 15 textual documents divided equally into three groups were created. Each group was first tested separately, then all the 15 documents at a time.

The recall is either 1.0 or 0.0 since there is only one base text to identify, while the precision was 0.066 for the 15 documents, and 0.2 for group test.

Repeating the process about 20 times by altering the texts in every cycle, it was found that 18 out of 20 repetitions were correct. The two cycles that could not identify the correct base text might be due to the length of the documents. In fact the CollateX library version 1.0 used in this study cannot currently handle large documents.

Knowing that the weighting system might not be convenient for every project, the possibility to modify the system was offered. In fact weights numbers have been for the time being arbitrarily selected.

This study is an attempt to convert a textual documents collation system into a document categorization one. It shows that by extending the Gothenburg model for basic collation system, it is possible to identify a base text and classify remaining documents.

7. Application

The system presented in this study could be used for several purposes, among them the followings.

a. Assisting memorization of text

The presented system possesses two editors that offer possibilities to interactively compare two texts side by side. A User could check his capacity of memorizing texts letter by letter through typing and visualization of results obtained from the collation process. Particularly people who memorize religious texts such as the Quran, which needs to be daily recited repeatedly, could considerably benefit from this system.

b. Plagiarism detection

The ranking process used in the system could be used to eliminate in advance documents that do not share some given percentage of proximity with a selected base text. Then qualified documents could be visually and interactively compared. A witness closer to a base text could be considered as a plagiat of the base text.

8. Conclusion and Scope of the Future

In this study, a textual document collation system using a rule-based Bayesian classifier for selecting a base text among several witnesses, and grading the remaining documents according to normalized frequencies of textual alterations found in those documents has been discussed.
A rule-based weighting system for normalizing frequencies of types of alterations has also been presented.
This study is a demonstration of using computational method for combining techniques of eclecticism, stemmatics, and copy-text, in order to identify a base text and grade remaining witnesses.

First eclecticism was used to identify types of textual alterations, then stemmatics was employed to cluster the witnesses based on the result of eclecticism, and finally copy-text was derived from the latter step.

Future work could encompass using other classifiers such as SVM to further validate the technique developed in this study.

Moreover the next paper will involve performing formal and quantitative empirical study on effects of types of textual alterations on the process of reconstructing lost original texts, in order to devise a more robust and mathematically convincing weighting system.

Enhancing the functionality of CollateX library by optimizing its processing time and runtime storage capacity, so that large documents could be properly collated, is also part of the tasks ahead.

References

[1] Mohamadou Nassourou, “Design and Implementation of Textual Documents Collation Systems”, accepted (poster presentation) in International Conference on Software, Services and Semantic Technologies (S3T 2011), Bourgas, Bulgaria, 2011, http://nbn-resolving.de/urn:nbn:de:bvb:20-opus-56601

[2] http://wiki.tei-c.org/index.php/Textual_alterations

[3] Theoretical Aspects of Textual Variations : http://www.bordalejo.net/DMU/ChapterIV.pdf

[4] G. Silva, H. Love, “The Identification Of Text Variants By Computer”, Inform. Stor. Retr. Vol. 5, pp. 89-108. Pergamon Press 1969

[5] Zhang J. (2004).”The Optimality of Naïve Bayes “, Proceedings of the 17th International Florida Artificial Intelligence Research Society Conferences, Florida, pp.562-567

[6] Manar Alkhatib Zayed, “Classification of Al-Hadith Al-Shareef Using Data Mining Algorithm”, University College of Information Technology. Abu Dhabi, UAE

[7] Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R. (2008),"Text Document pre-processing with the Bayes formula for Classification using the Support Vector Machine", IEEE Transactions on Knowledge and Data engineering, Volume 20, Issue 9, pp. 1264 – 1272

[8] Hirotoshi Taira “Text Categorization using Machine Learning”, Doctor’s Thesis, Nara Institute of Science and Technology, 2002

[9] http://en.wikipedia.org/wiki/Textual_criticism

[10] http://collatex.sourceforge.net/

[11] Zhang J. (2004).”The Optimality of Naïve Bayes “, Proceedings of the 17th International Florida Artificial Intelligence Research Society Conferences, Florida, pp.562-567

[12] Mohamadou Nassourou, “A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran”, accepted in WEBIST/WTM 2011, The Netherlands http://nbn-resolving.de/urn:nbn:de:bvb:20-opus-54712

[13] J. Price, “A Computer Aid for Textual Criticism”, Grace Theological Journal 8.1 (1987) 115-129.