Authorship Attribution based on Data Compression for Telugu Text

S. Nagaprasad
Research Scholar
Department of CSE
Aacharya Nagarjuna University
Guntur

P. Vijayapal Reddy, Ph.D.
Professor
Department of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad

A. Vinaya Babu, Ph.D.
Professor
Department of CSE
JNTU College of Engineering
Hyderabad

ABSTRACT

Authorship attribution (AA) can be defined as the task of inferring characteristics of a document's author from the textual characteristics of the document itself. In this paper we evaluated the compression model for AA on Telugu text. We considered six different compressors namely Zip, BZip, GZip, LZW, PPM and PPMd in combination with three different compression distance measures such as Normalized Compressor Distance (NCD), Compression Dissimilarity Measure (CDM) and Conditional Complexity of Compression (CCC). The result shows that the compression models are good alternatives for Authorship attribution instead of classification model with various features.

Keywords

Authorship attribution, Compressors, Compression distance measures, Macro-average, Micro-average, Accuracy, Telugu data set.

1. INTRODUCTION

Authorship attribution (AA) is a process of attributing unknown text documents to the correct known author from a set of known author set. Style characteristics of the author which are not under conscious control need to extract for authorship attribution. As the anonymous information increases in the web, research in authorship attribution becomes more important. The problem of authorship attribution is different from text categorization problem as writing style also needs to be considered in addition to text content which makes authorship attribution is a challenging task compared with text classification. For text categorization it is sufficient to consider only text content. Authorship attribution has many application areas such as a statement allegedly made by an accused person, plagiarism, forensic linguistics uses linguistic and statistical means, text alteration and authorship [22].

Authorship attribution research can be broadly categorized in two ways. Using a set of features and machine learning algorithms where as second way is by using compression algorithms. From the last decade, compression algorithms were effectively applied to group different text documents written by various authors [2]. In compression algorithms, compression distance between training and test document shows the similarity between two documents. A small distance between documents shows that test document is similar to the training set of an author, while a large distance indicates dissimilarity between test and training set. Data compression algorithms are best alternative approach for authorship attribution compared with the text classification model. File compression algorithms find duplicated strings in the text and checks for the longest matching strings. More frequent text sequences are coded with less byte where as rare sequences will be coded with more bytes [5].

The advantages data compression algorithms [23] for authorship attribution compared with classification model is that it avoids the word ambiguities, it considers only phrasal effects other than word boundaries, it deals with different types of documents uniformly.

The problem of authorship attribution for Telugu language text has not attempted. Various compression models are attempted on different languages text, but not on Telugu text. Hence it is required to be thoroughly test the influence of various compressors in combination with different distance measures on Telugu text for authorship attribution. In this paper an attempt is made on Telugu text for authorship attribution using different compressors with the combination of different distance measures.

2. RELATED WORK

There is an extensive research carried on authorship attribution using various features and with various classifiers. In [1] word length is used as a feature for authorship attribution. In [17] sentence lengths are used to judge authorship. The function word for authorship attribution is considered in [25]. The authors in [8] conducted experiments with support vector machine classifiers with various features. In [4] the study for authorship recognition implements multiple regressions and discriminated analysis. In [7] a function is generated to co-relate the word frequency and the text length. Karlgren-Cutting in [15] considered various style markers of the text for authorship attribution. Biber in [6] considered the syntactic and lexical style markers. Burrows in [3, 13] used principal components analysis (PCA) to combine various style markers which can discriminate among set of authors. In [15] machine learning algorithms such as naive bayes, decision tree and support vector machines were used to design discrimination models on large number of documents and features. In [17] author considered syllables per word using ngrams for authorship attribution. Stylometric features such as vocabulary richness and lexical repetition based on Zipf's [18] were studied on word frequency. Features such as word class frequencies, syntactic analysis, word collocations, grammatical errors, and word, sentence, clause, and paragraph lengths for authorship attribution were applied in [16].

As in [3] authorship attribution approaches can be distinguished based on how text documents are considered. Considering each training text individually known as instance-based approach or cumulatively known as profile-based approach. From the literature it was proved that most of the approaches reported follow the profile-based methodology [8].

Compression based authorship attribution is a new approach when compared to feature based authorship attribution. Compression algorithms are used to compress test documents and compare these compressed test documents with author profiles which is author wise compressed training documents sets. A high compression rate of test document with a particular author profile shows attribution towards that particular author. Many compression approaches were proposed for authorship attribution to assign test documents to corresponding author [5, 6, 7, 8]. Compression rate between documents, compression distances and other approaches are used to attribute a text. Preprocessing is not required for input documents while using compression algorithms for authorship attribution. Many compression methods have been used to attribute and categorize texts such as LZ76, LZ77, LZW, RAR, gzip, PPM. The method proposed in [7], shows good results with LZ76 where as other methods supports PPM family over LZ variants [9].

The compression algorithms build a dictionary or a model using trading text documents set. These generated models are used to the train classifiers. Test document can be assigned to a particular author by compressing this test document for each author specific model or dictionary which is generated during training phase. The test document is attributed to an author which is produced the highest compression rate [2]. In order to measure the compressed distance similarity many metrics were proposed in the literature [5, 29, 27].

Compression is the process of encoding original document using few numbers of bits. The process of authorship attribution using data compression is as follows. Given an unknown document d_i, and a set of training documents A_j of author j then compression algorithm S is applied to the original document set A_j and also to the concatenated of documents A_j and d_i such as $A_j + d_i$. The relative size after compression ΔS is then calculated as $S(A_j+d_i) - S(A_j)$ where $S(A_j+d_i)$ is the size of concatenated document after compression and $S(A_j)$ is the size of A_j after compression. The test document d_i is assigned to the author j if the smallest ΔS is computed with A_j. This difference is the cross-entropy between the two text documents.

Section 3 describes the methodology adopted for author identification using data compression, a brief description about various compressors and distance measures. The section 4 contains the experimental results and detailed discussion on the obtained results. The conclusions drawn from the discussions and possible feature extensions are mentioned in section 5.

3. PROPOSED APPROACH

Let C be a set of n authors, L is a set of training documents of all the known authors and T is a set of test documents. Authorship attribution method assigns each document from test set T to a candidate author from set C. In the first step, all the training texts of each author are concatenated and saved in one file. Concatenated training document per author is compressed using any compression algorithm which results to a author's profile, represents author's style. In the second step, the similarity between compressed test document t and the author profiles is computed. Then, the test document is assigned to one of the authors that minimize the similarity distance as shown in Figure 3.1.

3.1. Flow Chart

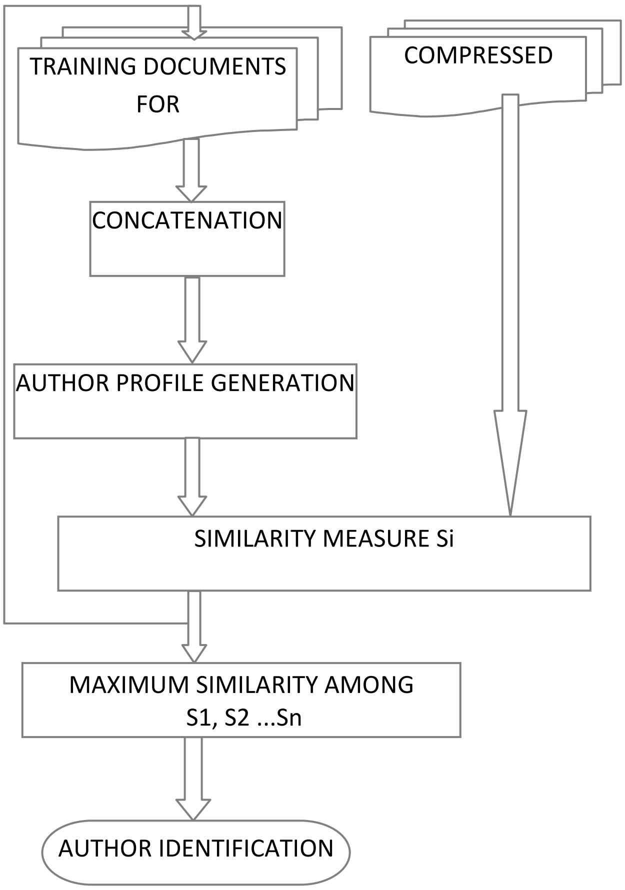

Figure 3.1: Flowchart for profile-based Authorship Attribution using data compression

3.2 Compressors

Data compression is the process of reducing the size of the data file. Compressors are to find the shortest sequence of bits needed to represent a text. There are two ways of compressing data namely lossless data compression and lossy data compression. For authorship attribution lossless data compression techniques are used. All data compression algorithms consist of two parts, a model which estimates the probability distribution and a coder which assigns the shortest codes to the most likely character.

The **Zip** compressor is a dictionary-based compressor used to compress text documents [9]. Text is compressed by verifying information that is repeated along the document and then represents it with a reference to the previously observed information. The **Bzip** compressor uses the Burrows-Wheeler algorithm with compression is done in blocks [26]. It compresses data in blocks of size between 100 and 900Kb using Burrows–Wheeler transform (BTW) block sorting text compression algorithm and Huffman coding. **GZip** is a dictionary based compression algorithm and uses a sliding window of 32 Kb to build the dictionary. If a training text is long enough then the beginning of that document will be ignored when GZIp attempts to compress the concatenation of that file with the unknown text. In Lempel Ziv Welch algorithm (**LZW**) compression algorithm the input file is read character by character and they are combined to form a string. The process continues till it reaches the end of file. Every new string is assigned some code and stored in Code table. They can be referred when the string is repeated with that code. Prediction by partial matching (**PPM**) is an adaptive finite-context method for text compression that is a back-off smoothing technique for finite-order Markov models [2, 18]. It obtains all information from the original data without feature engineering. **PPMd** is implemented by Shkarin [16] which is based on the basic PPM [17]. It uses a complex secondary escape estimation model and considers three cases such as binary context, nm-context and m-context.

3.3 Compression Distance Measures

Compression distances measures are used to compute a distance between two compressed text files. Compression based measures are used to estimate the amount of information shared by any two text documents. They can be utilized for clustering and classification on different types of data such as texts and images [20, 1].

Compression Dissimilarity Measure (CDM)

Compression Dissimilarity Measure (CDM) proposed in [1]. For documents x and y, the compression dissimilarity measure is defined as:

$$CDM(x,y)=\frac{C(xy)}{C(x)+C(y)} \quad (1)$$

Where C (x) is the size of the compressed object x, C (y) is the size of the compressed object y, xy is the concatenation of x and y and C (xy) is the size of the compressed object xy.

Normalized Compressor Distance (NCD)

Normalized Compression Distance (NCD) proposed in [2] uses general compressors to estimate the amount of shared information between two objects. The Normalized Compression Distance is defined as:

$$NCD(x,y)=\frac{C(xy)-min(C(x),C(y))}{max(C(x),C(y))} \quad (2)$$

Where C(x) is the size of the compressed object x. If x = y, the NCD is approximately 0, as the full string y can be described in terms of previous strings found in x; if x and y share no common information the NCD is 1 + e, where e is a small quantity due to imperfections characterizing real compressors.

Conditional Complexity of Compression (CCC)

Conditional Complexity of Compression (CCC) proposed in [5, 27]. The CCC of text y given text x is calculated by

$$CCC(y/x)=|S_c|-|x_c| \quad (3)$$

where $|x_c|$ is the length of the compressed text x. The S is the concatenated text of xy. CCC approximates a more abstract Kolmogorov conditional complexity and measures adapts to patterns in the training text for better compressing the unknown text.

3.4 Characteristics of Telugu language

The research on AA for Telugu language text has not attempted. There is a need to study the problem of Author identification as Indian languages are very rich in inflectional morphology [28]. Dravidian languages such as Telugu and Kannada are morphologically more complex compared with many languages in the world. Various compressors in combination with different distance measures are used on different languages text for Authorship identification, but not on Telugu text. Hence it is required to be thoroughly test the influence of different compressors with distance measures on Telugu text for AA. In this paper an attempt is made on Telugu text for AA using various compressors and distance measures with their combination.

4. RESULTS AND DISCUSSIONS

4.1 Description of Data Set

The dataset contains 300 Telugu news articles written by 12 authors which were collected from the Telugu News articles. The training set contains 20 documents per author where as testing set contains 5 documents per author. The training set is used to create the author profile for each author. The author profile is used to identify the author for each document from the test set.

4.2 Evaluation Metrics

In order to evaluate the performance of the proposed model, standard information retrieval metrics such as precision, recall and F1 -measure are used. Precision P_A , for author A, is defined as:

$$P_A=\frac{correct(A)}{retrieveddocuments(A)}=\frac{TP_A}{TP_A+FP_A} \quad (4)$$

Where TP_A (True Positive) is the number of documents that are correctly attributed to author A and FP_A (False Positive) is the number of documents that are incorrectly attributed to author A.

Recall R_A , for author A, is defined as:

$$R_A=\frac{correct(A)}{relevantdocuments(A)}=\frac{TP_A}{TP_A+FN_A} \quad (5)$$

Where FN_A (False Negative) is the number of missed attributions for author A.

F1 -measure, which is defined as the harmonic mean of recall and precision as:

$$F_1=\frac{2*P_A*R_A}{P_A+R_A} \quad (6)$$

F1 depends on author A. In order to aggregate these measures over all different authors' micro-average and macro-average were defined as follows.

Given a metric M (precision, recall or F1), for a set of n authors, these measures are defined as:

$$macro-average_M=1/n\sum_{i=1}^{n}M_{Ai} \quad (7)$$

$$micro-average_M=1/k\sum_{i=1}^{n}|D_{Ai}|M_{Ai} \quad (8)$$

Where k is the total number of test documents and $|D_{Ai}|$ is the number of documents in the test set for author A_i .

Accuracy is another measure and defined as:

$$Accuracy=\frac{\begin{array}{c}Numberofdocumentsthatare\\ correctlyassigned\end{array}}{Totalnumberoftestdocuments} \quad (9)$$

Experiments using the data set with various compressors in combination with distance measures were conducted. All the documents are compressed together to create the author's profile. The similarity is then computed between the compressed test document and the compressed author specific documents that contain author profiles. The obtained results are presented in Table 4.1, 4.2 and 4.3.

Table 4.1: Macro, Micro-F1 measures and accuracy for NCD for various compressors

Distance measure	**Normalized Compressor Distance (NCD)**		
Measure	**Macro_average F1 measure**	**Micro_average F1 measure**	**Accuracy**
Compressor			
ZIP	**0.55**	**0.55**	**0.72**
BZIP	**0.57**	**0.56**	**0.75**
GZIP	**0.62**	**0.61**	**0.78**
LWZ	**0.55**	**0.50**	**0.68**
PPM	**0.70**	**0.67**	**0.81**
PPMd	**0.74**	**0.74**	**0.85**

Table 4.2: Macro, Micro-F1 measures and accuracy for CDM for various compressors

Distance measure	**Compression Dissimilarity Measure (CDM)**		
Measure	**Macro_average F1 measure**	**Micro average F1 measure**	**Accuracy**
Compressor			
ZIP	**0.52**	**0.51**	**0.68**
BZIP	**0.55**	**0.55**	**0.72**
GZIP	**0.58**	**0.58**	**0.75**
LWZ	**0.51**	**0.51**	**0.66**
PPM	**0.68**	**0.62**	**0.78**
PPMd	**0.72**	**0.71**	**0.80**

Table 4.3: Macro, Micro-F1 measures and accuracy for CCC for various compressors

Distance measure	**Conditional Complexity of Compression (CCC)**		
Measure	**Macro_average F1 measure**	**Micro_average F1 measure**	**Accuracy**
Compressor			
ZIP	**0.58**	**0.57**	**0.75**
BZIP	**0.62**	**0.59**	**0.79**
GZIP	**0.65**	**0.65**	**0.82**
LWZ	**0.57**	**0.51**	**0.71**
PPM	**0.72**	**0.69**	**0.87**
PPMd	**0.78**	**0.76**	**0.89**

The six compression methods Zip, BZip, GZip, LWZ, PPM, PPMd and three distance measures CDM, NCD and CCC are used to test the performance. From the obtained results we can observe that PPM families of compressors are performing well with three distance measures compared with all other compressors. PPMd is much better compared with PPM in all three possible measures. PPMd is out performing in combination with Conditional Complexity of Compression (CCC) distance measure.

After PPM family of compressors Zip family of compressors are performing good. Among three Zip compressors such as Zip, BZip and GZip, GZip compressor's performance is good compared with remaining two Zip compressors with all three distance measures. BZip compressor's performance is less compared with GZip but better than Zip compressor. GZip compressor is performing well with Conditional Complexity of Compression (CCC) distance measure. Lempel Ziv Welch algorithm (LZW). Performance is worst when compared with all five remaining compressors in all three distance measures.

5. CONCLUSIONS AND FUTURE SCOPE

This paper evaluates the performance of compression-based similarity measures on authorship analysis on natural texts. There is no need to preselect which characteristics will be considered to classify the documents, since the classification is based on the similarity of those documents, measured by a normalized distance. In this study we have selected six different types of compressors: Zip, GZip, BZip, LWZ, PPM and PPMd. In order to compute the similarity between Author profile and test document, three different compression-based similarity measures were used in the experiments such as NCD (Normalized Compression Distance) and CCC (Conditional Complexity of Compression) and CDM (Compression Dissimilarity Measure). These 18 possible combinations were tested using profile-based attribution methods. Besides, our experimental results, it also show that the compression algorithms are an interesting alternative for authorship identification comparing favorably to traditional strategies based on feature extraction and classification. CCC seems more suitable for the profile-based approach compared with NCD and CDM.

In future work other compressors can be tested to verify if the authorship attribution with NCD can have better results with compressors that have a better compression ratio for text documents.

6. REFERENCES

[1] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 206–215.

[2] R. Cilibrasi and P. M. B. Vit, "Clustering by Compression," IEEE Transactions on Information Theory, vol. 51, no. 4, pp. 1523–1545, 2005.

[3] C. J. V. Rijsbergen, Information Retrieval, 2nd ed. Butterworth-Heinemann, 1979.

[4] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval, vol. 1, no. 1-2, pp. 69–90, 1999.

[5] D. Benedetto, E. Caglioti, and V. Loreto, "Language Trees and Zipping," Physical Review Letters, vol. 88, p. 048702, 2002.

[6] D. V. Khmelev and W. J. Teahan, "A repetition based measure for verification of text collections and for text categorization," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '03. New York, NY, USA: ACM, 2003, pp. 104–110.

[7] M. Lambers and C. Veenman, "Forensic Authorship Attribution Using Compression Distances to Prototypes," in Proceedings of the Third International Workshop on Computational Forensics. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 13–24.

[8] Y. Marton, N. Wu, and L. Hellerstein, "On compression-based text classification," In Proceedings of the European Conference on Information Retrieval, pp. 300–314, 2005.

[9] D. Sculley and C. E. Brodley, "Compression and Machine Learning: A New Perspective on Feature Space Vectors," in Proceedings of the Data Compression Conference, ser. DCC '06. Washington, DC, USA: IEEE Computer Society, 2006, p. 332.

[10] V. Bobicev, "Text Classification Using Word-Based PPM Models," The Computer Science Journal of Moldova, vol. 14, no. 2, pp. 183–201, 2006. 11. E. Frank, C. Chui, and I. H. Witten, "Text Categorization Using Compression Models," in Proceedings of the Conference on Data Compression, ser. DCC '00. Washington, DC, USA: IEEE Computer Society, 2000, p. 555.

[11] F. Peng, D. Schuurmans, and S. Wang, "Language and task independent text categorization with simple language models," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, ser. NAACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 110–117.

[12] C. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, pp. 379–423 & 623–656, 1948.

[13] A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," IEEE Transactions on Information Theory, vol. 22, no. 1, pp. 75–81, 1976.

[14] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," IEEE Transactions on Communications, vol. 32, no. 4, pp. 396–402, 1984.

[15] "PPM: One Step to Practicality," in Proceedings of the Data Compression Conference, ser. DCC '02. Washington, DC, USA: IEEE Computer Society, 2002, p. 202.

[16] P. G. Howard, "The Design and Analysis of Efficient Lossless Data Compression Systems," Providence, RI, USA, Tech. Rep., 1993.

[17] Bratko, A., Filipic, B.: Spam Filtering Using Compression Models. Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia, IJS-DP-9227 (2005)

[18] Cerra, D., Datcu, M., 2012. A fast compression-based similarity measure with applications to content-based image retrieval. Journal of Visual Communication and Image Representation 23 (2), 293 – 302.

[19] Watanabe, T., Sugawara, K., Sugihara, H., 2002. A new pattern representation scheme using data compression. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5), 579–590.

[20] Diederich, J., J. Kindermann, E. Leopold & G. Paass (2000) Authorship Attribution with Support Vector Machines. Applied Intelligence, 19(1-2), pp. 109-123.

[21] Forensic Linguistics Institute (FLI): http://www.thetext.co.uk/info.html.

[22] Frank, E., C. Chui & I. Witten (2000) Text Categorization Using Compression Models. Proceedings of the Data Compression Conference.

[23] Khmelev, D. & F. Tweedie (2001) Using Markov Chains for Identification of Writers. Literary and Linguistic Computing, 16(4), pp. 299-307.

[24] 25.S. Argamon, M. Sari ́, and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: First results. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 475–480, Washington, D.C., USA, 2003. ACM Press.

[25] 26. M. Burrows, D.J. Wheeler, A block-sorting lossless data compression algorithm. Technical Report 124, Digital SRC Research, 1994.

[26] 27. M. Malyutov, Authorship attribution of texts: a review, Electron. Notes Discrete Math. 21 (August) (2005) 353–357.

[27] 28.B.VishnuVardhan,P.VijaypalReddy, A.Govardhan"Corpus based Extractive summarization for Indic script", International Conference on Asian Language Processing (IALP) IEEE Computer Society (IALP 2011) pp 154-157

[28] 29. M. Li, X. Chen, X. Li, B. Ma, P. Vitanyi, The similarity metric, IEEE Trans. Inf. Theory 50 (December (12)) (2004) 3250–3264.