Determining the Relationship Between the Letters in the Voynich Manuscript Splitting the Text into Parts

Esbolat Sapargali[1], Iskander Akhmetov[1,2(✉)], Alexandr Pak[1,2], and Alexander Gelbukh[3]

[1] Faculty of Information Technology, Kazakh-British Technical University, Almaty, Kazakhstan
e_sapargali@kbtu.kz
[2] Institute of Information and Computational Technologies, Pushkin Street 125, Almaty, Kazakhstan
i.akhmetov@ipic.kz
[3] Instituto Politécnico Nacional, CIC, Mexico City, Mexico
gelbukh@gelbukh.com
http://kbtu.kz
http://iict.kz

Abstract. The Voynich Manuscript is an illustrated manuscript code that has not yet been defined the structure of the writing and the relationship to other languages. This study investigated the effectiveness of examining point detail versus examining the full picture all at once in a single study. In the approach of this study, one of these ways, some letter patterns based on frequency and word length were identified, including connections at different combinations of consonant and vowel letters by a statistical approach for a hidden Markov model. A narrowly directed systematic direction can help lead to the unraveling of the manuscript text in progressive steps.

Keywords: Voynich manuscript · Hidden Markov model · Voynich manuscript characters · Natural language processing

1 Introduction

The Voynich manuscript is an illustrated handwritten code written in an unknown language or cipher. It is still one of the main unsolved problems of linguistics and cryptography. The problem of deciphering it is that scientists have not yet been able to determine the structure of the writing and the relationship to other languages.

Previous studies (the first attempts still date back to the 16th century) have tried to decipher the text by all known methods of cryptography and linguistics, including advanced technologies such as neural networks. But large-scale research has failed to make breakthrough advances in deciphering [9].

I. Batyrshin et al. (Eds.): MICAI 2021, LNAI 13068, pp. 163–170, 2021.
https://doi.org/10.1007/978-3-030-89820-5_13

The purpose of this article is to determine the effectiveness of studying point details, compared to studying the whole picture at once in a single study. The approach of this study will use methods that allow statistical analysis of words in the text. Here statistical methods will be used for the hidden Markov model. This will be used to find out to what extent there is a pattern in the various parameters of word frequency, word types, and the removal of spaces between words. The contribution will be that a narrowly directed systematic study will help to arrive at a clue to the manuscript text in small steps.

2 Literature Review

The exact time when the manuscript was made remains unknown, but its history can be traced back to the 16th century. The alphabet of the manuscript bears no visual resemblance to any known writing system, the text has not yet been deciphered, and illustrations (women's clothing and attire as well as a couple of castles) are used to determine the age of the book and its origin. Scholars have concluded that the details are characteristic of 15th- and 16th-century Europe [4].

In 1912 the Society of Jesus was in need of funds and decided to sell some of its property. While sorting through the coffers of books in the Villa Mondragone, Wilfried Voynich stumbled upon a mysterious manuscript, which to this day bears his name [8]. Throughout the century, various methods have been used to decipher the manuscript, and if we focus on them, we can see small steps toward discovery thanks to discoveries made by researchers. Since then, a number of claims have been made about possible decipherment and a number of hypotheses have been advanced, none of which, however, has been unequivocally confirmed and accepted in the scientific community. This literature review will focus on research on word analysis using statistical methods of analysis [3].

Reddy and Knight in their study did not use word-value for character sets divided into words. They did the opposite, dividing them into two types, lexemes and words. They even calculated the exact number of each. The following article says that single-character words occur. They are mostly characters similar to 2 and 9 [6].

In linguistic analysis, there is a lot of discussion about word statistics versus other statistics. There are transliteration issues in addition to word space definition. The most reliable of these is the word length statistic. In a related study, Jorge Stofi concluded that, according to his definition of symbols and spaces, the distribution of word lengths is binomial. This is an unusual phenomenon for natural language, which until now has not been understood or explained [7].

Of more recent research, we can mention Christian Perone, who, using word vectors and visualization of some t-SNE models, has shown how useful word vectors are for analyzing unknown codewords for grouping vector space, for example, for deriving platform names from a lower representation. A University of Adelaide research project led by Dewitt Abbott has developed a research scheme for decoding manuscripts. Including an important questionnaire vector method for statistical analysis [5].

It can be seen from previous studies that one can do the calculations and expect to get different results for the lexeme word length distribution and the word type length distribution, since the more frequent word types tend to be shorter. This is true for ordinary languages as well as for the Voynich text.

The Voynich manuscript was studied using a static Hidden Markov Model (HMM). The HMM is a generative probabilistic model in which the sequence of observed variables is generated by a sequence of internal hidden states. Transitions between latent states are assumed to have the form of a Markov chain. They can be defined by a vector of onset probability and a transition probability matrix a transition probability matrix to determine the probability of the sequence between letters. The outlier probability of the observable can be any distribution with parameters due to the current hidden state [2].

The HMM uses only the variables that are affected by a given state to investigate, so the transition probabilities are the only parameter. Each state has a probability distribution among all possible output values. Therefore, the sequence of generated symbols provides information about sequence of symbol(s) states [1].

3 Methodology

The Voynich manuscript was studied using a static Hidden Markov Model (HMM). In contrast to entropy and mutual information methods, The hidden Markov model can analyze the relations between letters without their prior segmentation, because the sequence of symbols can be extracted from the columns of the scanned symbol image in the same way as from the word image. Segments can then have character sequences (which are jointly optimized by the separation and classification result) extracted from their strings, and the string HMM can be effectively used to check the classification for each segment hypothesized by the column HMM [2].

The diagram below shows the general structure of the HMM. The ovals represent variables with a random value. The random variable x(t) represents the value of the latent variable at time t. The random variable y(t) is the value of the observed variable at time t. The arrows in the diagram symbolize conditional dependencies. It is clear from the diagram that the value of the latent variable x at time t depends only on the value of the latent variable x at time t−1. This is called the Markov property. Although, at the same time, the value of the observed variable y(t) depends only on the value of the latent variable x(t).

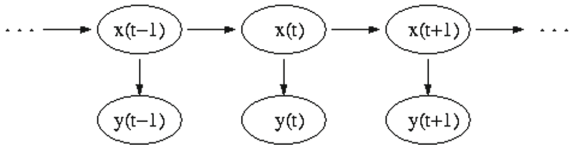

Fig. 1. Temporal evolution of a hidden Markov model

The project used scanned images of the manuscript as well as texts divided into different types of parts to determine the relationship of certain letters as well as substitutions for other characters. Functions were created for tokenization, word counts, and bigram concurrency. Some letter patterns based on word frequency and length have been identified, including relationships at different combinations of consonant and vowel letters.

During tokenization the text is an array. The function written for tokenization selects strings longer than 4 words. The input, of course, is the list of words we want to tokenize. Then a two-part HMM fitting takes place, and we get the tags. The tags themselves are highlighted in green, and all other occurrences are highlighted in blue.

Old manuscripts have folios in which words with the same meaning are arranged together to provide similarity queries. The data are sorted by several properties: by folio, by line, and by word length. Minimal pairs are pairs of words or phrases in a particular language, spoken or written, that differ in only one phonological element.

Word matches in sentences are called bigram frequencies. As you know, for bigrams, the context window is asymmetric by one word to the right of the current word when counting occurrences together. The morpheme boundary runs between the morphemes that make up the word. In some cases, the free stem and suffix are connected by a morpheme boundary, but in most cases the bases are also connected. In the first case, the free base, and in the second case, the connected base.

The "Editing Distance" part of the code calculates the Levenshtein editing distance between two strings. The edit distance refers to the number of characters that must be replaced, inserted, or deleted to convert s1 to s2 [6].

4 Experiment

HMMlearn implements Hidden Markov Models. Each HMM parameter has a symbolic code that can be used to configure its initialization and evaluation. An initial point is required for the EM algorithm to proceed; before training, each parameter is assigned a value, either random or computed from the data. The input data for training is a matrix of combined observation sequences along with sequence lengths.

The Voynich manuscript document is presented as extracted features, not as a line form. When preparing the datasets, The dataset was loaded with Voynich text from vdata. Initially, the text is divided into parts, writing how many words are in each part, which line of the paragraph/folio. A Pandas word dictionary has also been built. The dictionary will be used to preserve the temporal vocabulary and the manuscript word corpus. Only whitespace characters were used for tokenization and then each word was highlighted in lower case.

In addition to the text, there are also scanned images of the manuscript. They are taken from the voinichese website. They were used as a reference in

	folio	paragraph	line	text	words
0	f1r	P1	1	fachys ykal ar ataiin shol shory cthres y kor ...	10
1	f1r	P1	2	sory ckhar or y kair chtaiin shar are cthar ct...	11
2	f1r	P1	3	syaiir sheky or ykaiin shod cthoary cthes dara...	9

Fig. 2. The output of Voynich's adapted manuscript line by line

certain situations. Before examining the folio, its integrity was checked. This was done to be sure that we could select one paragraph at a time and combine them.

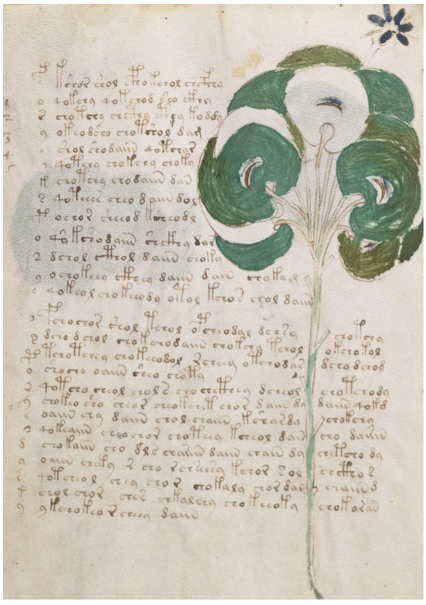

Fig. 3. Fragment of a scanned image of the manuscript

Next we move on to minimal pairs, Such as phoneme, toneme, or chroneme, and have different meanings. They are used to demonstrate that two phonemes are two separate phonemes in a language.

A sliding window was used to look at them, taking 21 sheets, averaging their number of lines, and plotting them on the entire manuscript. Each column is a line on a graph, and the height of the line is the average line length in that window.

Next, we calculate how many times a pair of words occur in sentences, regardless of their position in the sentences. We calculated only half of the matrix, since we only need the editing distance of the equivalent of 1.

5 Result

As a result of the study, the folio integrity check function showed a match after several check attempts. Searching for specific characters and finding a pair gave a successful result. The number of unique words was determined (8078 out of 37886). After that, we displayed a graph of frequently encountered pairs.

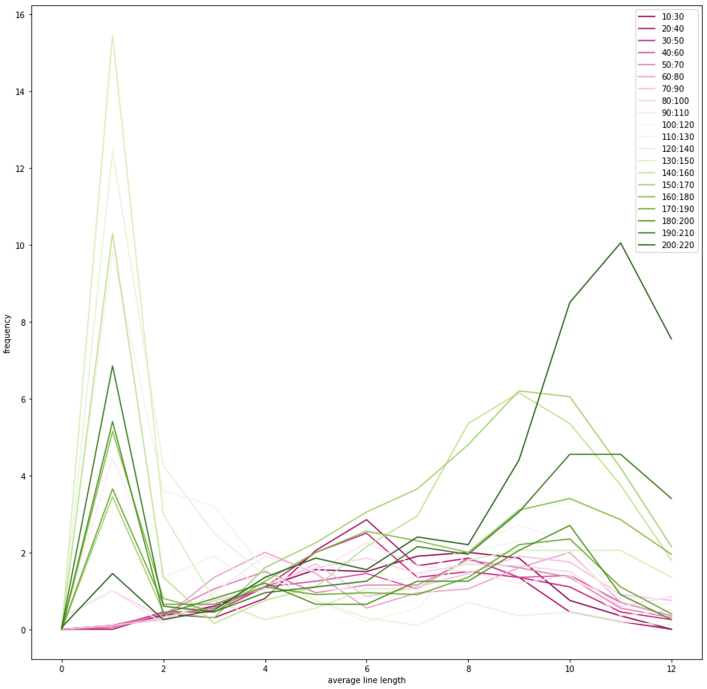

Fig. 4. Sliding window with 21 sheets averaged over the entire manuscript

With the help of the graph and table it was possible to understand the patterns with the help of visualization functions, looking at the pictures. The folio consists of an average of 5 lines of 5 words each (a long row of properties). All the words in the left margin come first.

The data sets are arranged in letter columns. This is interpreted as a 21-window counter capturing more details about the appearance of each word than other methods.

The configuration file generated the best choice of vowels and consonants (some groups of letters were vowels and some were consonants) showed V = ai, e, o, k, C = a, l, dy, n. Loss of vowels between consonants would be 0, with only consonants or only vowels would be 0.1. When in condition 1, the probability of seeing these words will be frequent. But 5.5% when going to state 2 other words.

There are letters that are different. Returns context, letters with state 1 or 2. No difference was detected on the first pass. The first insertion was returned, regardless of editing distance. Some letters (d) replace each other with a higher frequency than others. K t have relations, o replaces a, e and y in many cases.

Some groups of letters turned out to be vowels and some groups turned out to be consonants. The configuration file generated the best selection of vowels and consonants and showed V = ai, e, o, k, C = a, l, dy, n. The loss of vowels between consonants would be 0, with only consonants or vowels would be 0.1.

This could mean when being in condition 1, the probability would be frequent to see these words. But 5.5% when going to state 2 you will see other words.

Got the letters, which is different. Returns context, letters with state 1 or 2. No difference was found in the first pass. The first insertion was returned, regardless of editing distance. Some letters (d) are replaced with each other at a higher frequency than others. K/t have a relationship, o replaces a, e and y in many cases.

6 Conclusion

In this study, using documents, scanned images of manuscripts, as well as text divided into different types of parts, it was possible to determine the relationship of certain letters, as well as their replacement by other characters using the static Markov model and bigram. Scanned images of symbols and words from the manuscript as well as texts divided into different types of parts were examined.

Letter patterns (ratios of certain letters) from the frequency and length of words were determined, including the occurrence of certain letters in English as well as the occurrence of consonants and vowels. Some letters are substituted for each other with greater frequency than others and also have a substituting relationship in some cases.

Some letter patterns based on word frequency and length have been identified, including relationships at different combinations of consonant and vowel letters.

Compared to previous research, it has become clear that different features are required for different applications. The ways in which the Voynich manuscript is handled are varied and depend on applications and languages, which means that they cannot be restricted to any general framework in the subsequent study.

You can then develop a function appropriate to the input format (catalog browsing, eXtensible Markup Language parsing). To get a new list of tokens in each document, the input can be parsed, and the tokens can also be converted into folders to identifiers and displayed inside the repetition with the resulting sparse vector.

Acknowledgment. We gratefully acknowledge the financial support of the Ministry of Education and Sciences, Republic of Kazakhstan (Grant num. AP09260670 "Development of methods and algorithms for augmentation of input data for modifying vector embeddings of words")

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

References

1. Acedo, L.: A hidden Markov model for the linguistic analysis of the Voynich manuscript. Math. Comput. Appl. **24**, 14 (2019). https://doi.org/10.3390/mca24010014
2. Cave, R., Lee, N.: Hidden Markov models for English. In: Ferguson, J.D. (ed.) Hidden Markov Models for Speech. IDA-CRD, October 1980
3. D'Imperio, M.: An application of cluster analysis and multiple scaling to the question of "hands" and "languages" in the Voynich manuscript. Nat. Secur. Agency Tech. J. **23**, 59–75 (1978)
4. Levitov, L.: Solution of the Voynich Manuscript?: A Liturgical Manual for the Endura Rite of The Cathari Heresi. The Cult of Isis. Aegean Park Press, Laguna Hills (1987)
5. Perone, C.: Voynich Manuscript: word vectors and t-SNE visualization of some patterns (2016). https://blog.christianperone.com/2016/01/voynich-manuscript-word-vectors-and-t-sne-visualization-of-some-patterns/
6. Reddy, S., Knight, K.: What we know about the Voynich manuscript. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 78–86. Association for Computational Linguistics (2011)
7. Stolfi, J.: Voynich Manuscript stuff (2005). https://www.ic.unicamp.br/stolfi/voynich/
8. Voynich, W.: A Preliminary Sketch of the History of the Roger Bacon Cipher Manuscript, pp. 415–430. Transactions of the College of Physicians of Philadelphia, Printed by T. Dobson (1921). http://resource.nlm.nih.gov/2546054R
9. Zandbergen, R.: Voynich MS - history of research of the MS (2019). http://www.voynich.nu/solvers.html