Heritage Science

RESEARCH Open Access

Created era estimation of old Korean documents via deep neural network

Inseon Yoo[1] and Hyuntai Kim[2*]

Abstract

In general, the created era of a literary work is significant information for understanding the background and the literary interpretation of the work. However, in the case of literary works of old Korea, especially works created in Hangul, there are few works of which the era of creation are known. In this paper, the created era of old Korean documents was estimated based on artificial intelligence. Hangul, a Korean letter system where one syllable is one character, has more than 10,000 combinations of characters, so it is available to predict changes in the structure or grammar of Hangul by analyzing the frequency of characters. Accordingly, a deep neural network model was constructed based on the term frequency of each character in Hangul. Model training was performed based on 496 documents with known publication years, and the mean-absolute-error of the test set for the entire prediction range from 1447 to 1934 was 13.77 years for test sets and 15.8 years for validation sets, which is less than an error ratio of 3.25% compared to the total year range. In addition, the predicted results of works from which only the approximate creation time was inferred were also within the range, and the predicted creation years for other divisions of the identical novel were similar. These results show that the deep neural network model based on character term frequency predicted the creation era of old Korean documents properly. This study is expected to support the literary history of Korea within the period from 15C to 19C by predicting the period of creation or enjoyment of the work. In addition, the method and algorithm using syllable term frequency are believed to have the potential to apply in other language documents.

Keywords: Old Korean document, Hangul, Created era estimation, Term frequency, Deen neural network

Introduction

Because of increase of the computing power via nanotechnology and the development of algorithm which started from perceptron, the artificial intelligence (AI) research field has received enormous attention [1–5]. AI technology is applied in numerous fields, such as image recognition, engineering, and natural language processing (NLP) [6–10]. Especially, NLP which treats human language, has large potential to be applied to speech recognition, translation, text processing, and speech synthesis [10–13]. NLP is also applied to literature research field. Various attempts are being studied, such as classifying a long novel based on clustering, comparative studies of different versions of the identical work, and creating literary works through pattern learning [14–20].

In literature research, the created era of the work is an important milestone [18, 21, 22]. If the created period is known, a fertile study of the literature is available by explaining the characteristics of the work in relation to the created/enjoyed times. The timeline of different works is also important as one may predict the influence relationship of works. The created era also provides information about when the work was actively enjoyed.

The language system of the Chosen Dynasty was a bilingual system using both Chinese characters and Hangeul. However, the importance of Hangeul was very low compared to the Chinese character, which was the official letter. The upper-class men who led the Chosen Dynasty valued Chinese characters, used Chinese characters

*Correspondence: hyuntai@hongik.ac.kr

[2] Electrical and Electronic Convergence Department, Hongik University, Sejong, Korea
Full list of author information is available at the end of the article

for official political and literary activities, and avoided using Hangeul. On the other hand, women and lower classes, that is, those who had relatively few opportunities to learn and were marginalized from the mainstream, learned and used Hangul, because it is relatively easy to learn. From a functional point of view, Chinese characters were used for official records, and Hangeul was used only for personal letters and novels that were not properly valued at the time [23]. Accordingly, in the case of texts created in Chinese characters, the author and year of creation are usually clearly identified, whereas, in the case of texts created in Korean, the author and created year are often left unknown. Therefore, the work of estimating the crated age of Hangeul documents, which remained unknown for a fairly long period of time, about 500 years after Hangeul was promulgated in 1446, has an important meaning, as the work illuminates and understands aspects of culture at the time that were not properly recognized and remained unofficial. By estimating the created year of the Hangeul documents of the Chosen Dynasty, the literary trends and streams of the time can be grasped, and by understanding the social customs reflected in the texts, one can understand the aspect of society and culture. The work will be helpful in drawing a complete picture of the cultural topography of the Chosen Dynasty by allowing both the official culture represented by Chinese characters and the informal culture represented by Hangeul to be considered.

Especially in Korean literature, research based on data analysis and AI is now at a preliminary level. Created era estimation research of classical Korean documents is not studied intensively, to the best of our knowledge. In this paper, we propose a method based on deep neural network (DNN) to estimate the created era of Korean classical documents. The DNN model is trained based on the character term frequency (TF) of the document. [18, 24, 25] Each TF of the single Korean character is calculated via data processing, and the output of the model predicts the created era of the document.

Principle and methods

Structure of Korean character

Hangul is an interesting phonetic character. Similar to the alphabet, Hangul is a phonemic language, but a single Korean alphabet cannot exist on its own [26–28]. A Korean character is formed by a combination of first sound, middle sound, and final sound / or only first sound and middle sound. First sounds and last sounds are consonants, and middle sounds are vowels. One completed character sounds one syllable. For example, if one would like to transfer word "telephone" to Korean character, three syllables of 'tel', 'le', and 'phone' become each character. The first letter 'tel (tel)' is converted to '텔', where 't' sounds like 'ㅌ', 'e' sounds like 'ㅔ', and 'l' sounds like 'ㄹ'. The second letter 'le (lɪ)' is converted to '레', where 'l' sounds like 'ㄹ' and 'e (ɪ)' sounds like 'ㅔ'. Note that sound 'e' and 'ɪ'are difficult to be distinguished in Hangul system. It is similar to convert from Hangul to English. For example, a word '방탄' (bulletproof) is composed with two letters, which means it has two syllables. The first character '방 ' sounds like 'bang', and the second letter '탄' sounds like 'tan'. Details are shown in Fig. 1.

There are 14 basic consonants, 10 basic vowels, double consonants (example: ㄱ+ㄱ = ㄲ, ㅂ+ㅅ=ㅄ) and double vowels (ㅗ+ ㅣ=ㅚ, ㅓ+ ㅣ= ㅔ) in contemporary (modern) Hangul. The combination of these alphabets generates more than 10,000 combinations of characters. In terms of Hangle of old era, there were additional consonants such as ㅿ, ㆆ, double consonants which are not used on contemporary Hangul such as ㅼ, an additional vowel ㆍ, and double vowels which are not used in contemporary Hangul such as ㆉ. These old Korean characters have gradually disappeared or transformed.

As explained, there are numerous characters, or letters in Korean compared to alphabet. Various letters, more than 10,000, enables to perform document analysis based on character TF, which reflects the frequent word, structure, or the spelling system of Hangul of the era.

Data processing of old Korean documents

Contemporary Korean characters are defined in Unicode 44,032~55,203, and frequently used old Korean characters are distributed in Unicode 57532~63086. [29] In this paper, we perform the data analysis based on database "21st century Sejong Project", which has old Korean documents and corresponding dates of publication. For the training, we consider 560 documents with exact published year, and published before 1934. The old Korean when the old Hangul system was officially ended in 1933, however still there were uses in old Hangul spelling systems even in 1934 for some documents. [26, 30] We extract and calculate the TF of old Korean characters

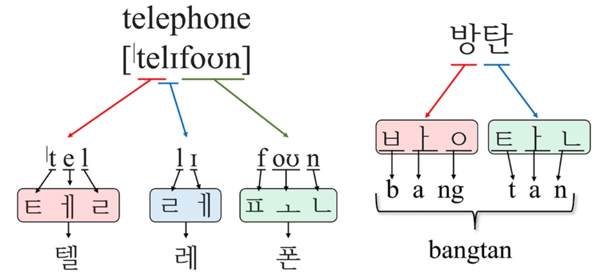

Fig. 1 Schematic of the transfer process between English and Hangul

and contemporary Korean characters. Assuming the total length as l and the TF as n, three kinds of term frequencies—raw TF (n/l), log scale TF ($log(1+n/l)$, and binary frequency ($sign(n)$, result to be 0 or 1)—has been considered.

Among the documents, there are works with a high proportion of Chinese characters, and works with a very small portion of old Korean characters. Documents with high Chinese character ratio are works in which only the propositions and postpositions are in Hangul, and documents without the old Hangul are materials typed in modern Korean character. These documents are not proper to estimate the created era, so documents with a high ratio of Chinese character and a low ratio of old Korean character have been removed from the training dataset. In addition, we have excluded characters with extremely small ratios. These infrequent characters could be a typo or letters with less importance.

For the input and output of the model, all three TFs are considered. Two different inputs are regarded, one with only the old Korean TFs, and the other is all TF with both old and contemporary Hangul. The output is the created era of the document. Hereinafter, the first model which considers old Korean is called old Korean model, and the other is called all Korean model. The year of the datasets are from 1447 to 1934, and they are normalized from 0 to 1.

Neural network model

Based on the TF data from the document, two DNN models are assumed for created era estimation, where one model only regards old Korean TFs as an input, and the other considers both old and contemporary Korean. Two hidden multilayer perceptrons (MLPs) are positioned between the output, and a single output that represents the created era is positioned at the end of the model. MLP layers are connected with rectified linear unit (ReLU) activation function, and 5% of dropout is added to prevent overfitting. The final layer is fully connected, and the activation function was chosen to be sigmoid function. The model summary is depicted in Fig. 2. The first MLP layer has 256 states, and the second MLP layer has 128 states. The optimizer has been selected to be 'ADAM', and root-mean-square (RMS) has been selected to be the loss (cost) function. Note that the RMS cost function is the square root of the mean-squared-error (MSE). The batch size of training is selected as 64. In terms of complexity, the old Korean model has 522,497 trainable parameters and the all Korean model has 1,457,921 trainable parameters, respectively.

Results

Training and test on documents with known year of creation

Based on the methods in the previous section, data extraction and DNN model training have been performed. Figure 3a shows the created year of the 560 documents. The document IDs are sorted in terms of created year. Extracting the Korean characters, 2220 kinds of old Korean characters and 3205 kinds of contemporary Korean characters were observed. Total of 5425 kind of Korean characters exists on 560 documents. The binary TFs of the characters on each document are depicted in Fig. 3b. The character ID from 1 to 2220 represents the old Korean (left side of the red dashed line), and it is observed that old Koreans tend to disappear while the document gets closer to modern times. Some disappeared characters are shown in the blue box.

From the 5425 characters, infrequent letters which has TF lower than 5×10^{-4} are removed, therefore 637 kinds of old Hangul and 1218 kinds of contemporary Hangul are considered. In terms of input, three TFs are considered, so for the old Korean case, $637 \times 3 = 1911$ TFs are considered as an input vector, while the case of

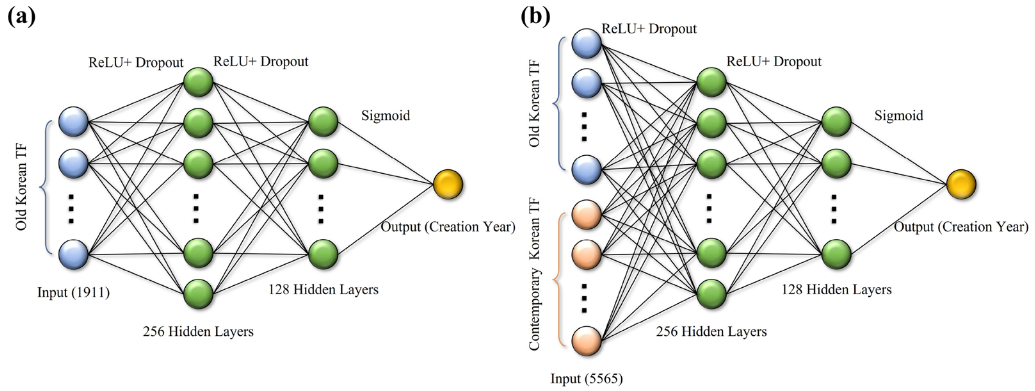

Fig. 2 The DNN architecture of model which considers (**a**) only old Korean and (**b**) all Korean

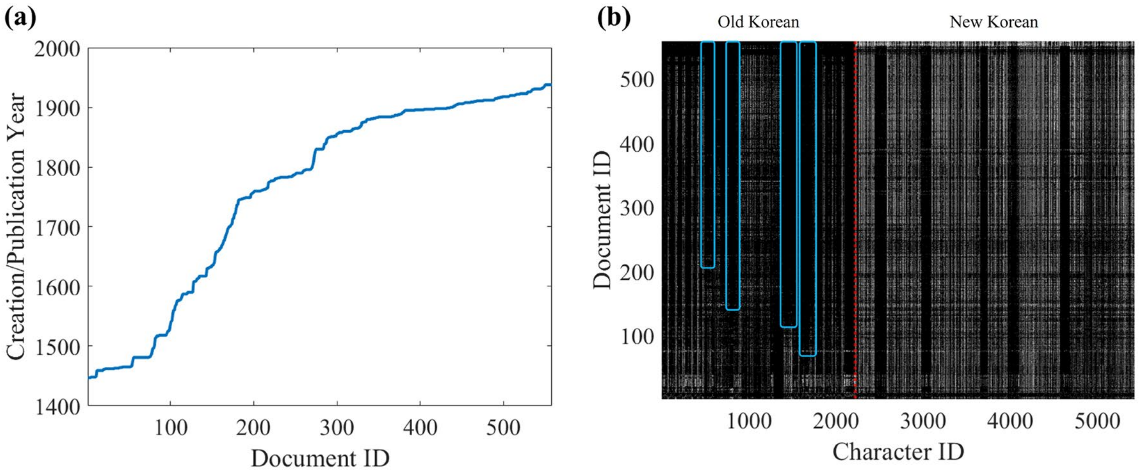

Fig. 3 **a** The document ID and publication year of the "21st century Sejong project" database. **b** Binary TF of the Korean characters in terms of document ID

all Korean, $(637 + 1218) \times 3 = 5565$ TFs are the input vector.

To remove documents that have a high proportion of Chinese character, the ratio of Korean letters within the total document compared to the total length including Chinese characters are calculated. To check the documents with a low old Korean ratio, the old Korean ratio compared to all Korean letters is also calculated. Both results are depicted in Fig. 4. It is shown that some documents have relatively low Korean ratios and old Korean ratios. Korean ratio under 40%—which are documents mainly written in Chinese characters—and old Korean ratio under 7.5%—which the old Korean system is deformed while typing to digital files—have been removed from the data to be considered. After removing the outliers, 496 sets of documents remained, and these are considered to be the documents to train the DNN model. From Fig. 4, it is observed that the old Korean ratio tends to be decreasing, because the old Hangul gradually disappears as we get closer to modern times.

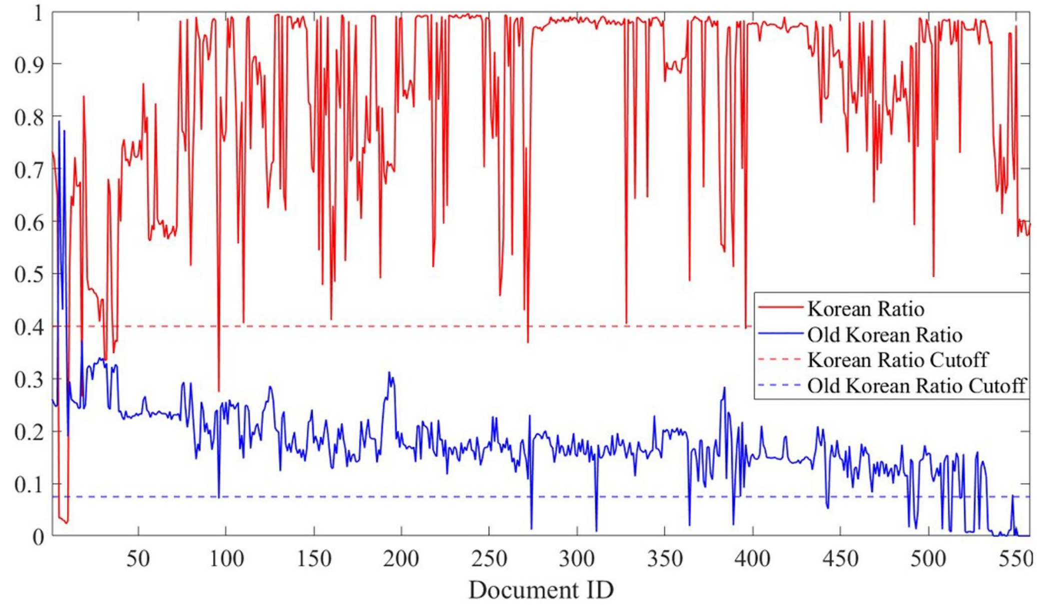

Fig. 4 The Korean ratio and old Korean ratio of the documents. The cutoff level for data purification is shown as dashed lines

The final 496 sets have been divided into train-test-validation sets of a ratio as 79.8%, 10.1%, and 10.1%, which results 50 test sets and 50 validation sets. The RMS loss of training sets and validation sets within 300 epochs are depicted in Fig. 5a. Note that the loss has been rescaled to year unit. The number of data is limited, so the model cannot predict the validation set with high accuracy, however, the error is shown to be converged on a certain level. Figure 5b shows the predicted year via both models in terms of actual publication year. Some data shows large errors, however, most of the data shows similar trend compared to the actual year.

The RMS loss and mean-absolute-error (MAE) error of the validation set and test are calculated and shown in Table 1. From both results, it is notable that the model considering all Koreans has higher accuracy, because considering all Koreans have more inputs to consider. In addition, the change of language and Korean system are not only reflected in old Korean, but also in contemporary Korean. The total estimation range is 487 years which is from 1447 to 1934, the results show at most 4.7% of RMS error and 3.24% of MAE for the all Korean model. In terms of MAE, the first quartile was 4.13 for old Korean model and 3.43 for all Korean model. The median MAE was 12.92 for old Korean model, and 9.4 for all Korean model. The third quartile was calculated as 27.29 for old Korean model and 18.37 for all Korean model. The ratios of loss less than 50 were 91% in the old Korean model and 96% in the all Korean model. It is worth noting that there are fewer data during 1600∼1700, therefore the accuracy near 17th century tend to be low.

Prediction for unknown documents

Among the data in "21st century Sejong Project", there are documents whose publication year is not defined. Only the estimated created century is listed for some of the documents, and most of the publication year of the documents is unknown. We apply our model to documents whose publication century is estimated roughly. Documents which is estimated to be publicized in 16th century are "救急簡易方諺解", "구급간이방언해", and "순천김씨묘출토간찰". "두창경험방", "셔궁일긔", and "진주하씨묘출토간찰" are believed to have been published in the 17th century. "聘聘傳" and "응진경언해 " are expected to be published in the 18th century. and "廣才物譜" is estimated to be a document of 19th century. The predicted year of publication by two models are shown in Table 2. Considering ∼20 year of error range, most of the documents are within the previously estimated century.

Table 1 The RMS and MAE loss of old Korean model and all Korean model

(Year)	Old Korean model		All Korean model	
	RMS	MAE	RMS	MAE
Test	29.46	16.97	22.87	13.77
Validation	31.19	22.36	22.07	15.80

In "21st century Sejong Project" database, there are "Gasa" (Korean poem) collection of 17th and 18th centuries which has been published in 20th century. The predicted year of these Gasa collections are shown in Table 3. As the publication year is 20C, some of the expressions would confuse the model, therefore the predicted created year will be slightly increased. Most of the 17C documents are predicted to be 18C documents, and 18C documents to be around 18C or 19C.

We have also predicted the publication year of Korean classical novels, "엄씨효문청행록", "명주보월빙, "옥루몽 ", and "소현성록". [18] The documents are known to be

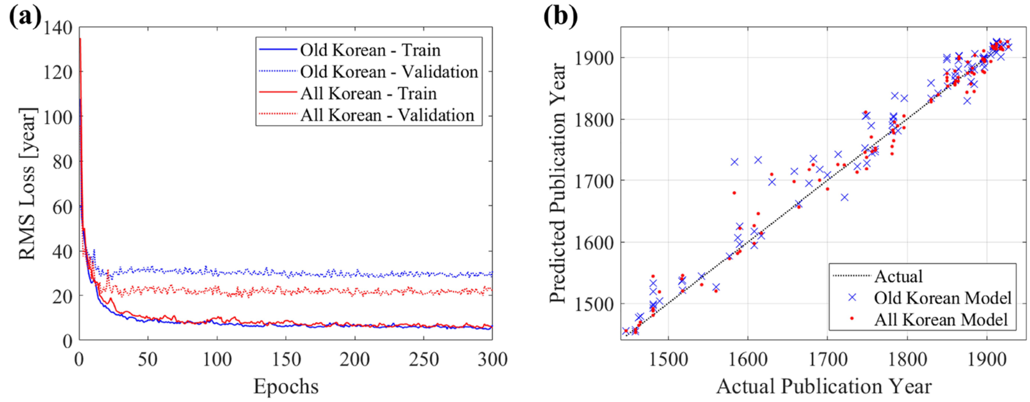

Fig. 5 **a** RMS loss of two models while training. **b** Predicted year from the two models for the test and validation sets

Table 2 Prediction of documents for which the approximate era of creation is known

Estimated era	Title	Old Korean model	All Korean model
16C	救急簡易方諺解	1498	1506
	구급간이방언해	1498	1506
	순천김씨묘출토간찰	1632	1635
17C	두창경험방	1726	1739
	셔궁일긔	1777	1779
	진주하씨묘출토간찰	1671	1699
18C	聘聘傳	1795	1770
	응진경언해	1843	1890

Table 3 Prediction of Gasa collections

Estimated era	Title	Old Korean model	All Korean model
17C	立巖別曲	1712	1733
	農夫詞	1702	1673
	時歎詞	1717	1713
	所有亭歌	1712	1724
	白馬江歌	1854	1795
	採薇歌	1726	1672
18C	기성별곡	1831	1826
	단산별곡	1805	1773
	금강별곡	1920	1899
	마천별곡	1689	1768

created between 18C and 19C. As the volumes of the novels are vast, the database divides the novel into several pieces. The predicted results are shown in Table 4.

The results show that predicted year is within the range of the known created era. In addition, for identical novels, different partitions show considerably similar predicted publication year. As the overall known created time and the calculated estimated year are similar, the validity of the predictive model of this study is considered to be somewhat reliable.

Discussion

There are several steps when a document is transferred to digitized data. The created original version is copied several times by hand copy or printing. The published version could be the original version or the copied edition. Variations in Korean character system, grammar, or expression will be reflected in the publication version. In addition, extra variant can be applied while typing the published version to digital document. The process from creation to digitization is shown in Fig. 6.

This explains why the 'Gasa' collections, which are publicized in 19C but the contents created in 17C or 18C, are predicted to be created/publicized later than the created year. However, even the publicized version is varied on various steps, the publication year still is a valuable material, as it informs the minimum limit of the created year.

The old Korean model and all Korean model shows different results, and all Korean model is shown to have slightly better performance. The changes in system and

Table 4 Prediction of Korean classical novels

Novel title	Division	Old Korean model	All Korean model
엄씨효문청행록 (장서각본)	Part 1	1889	1872
	Part 2	1838	1835
	Part 3	1837	1846
	Part 4	1840	1827
	Part 5	1859	1863
명주보월빙 (장서각본)	Part 1	1816	1805
	Part 2	1823	1805
	Part 3	1822	1823
	Part 4	1806	1804
	Part 5	1820	1818
	Part 6	1811	1806
	Part 7	1842	1826
	Part 8	1808	1807
	Part 9	1797	1794
	Part 10	1808	1796
	Part 11	1820	1808
	Part 12	1826	1810
	Part 13	1822	1830
	Part 14	1804	1802
	Part 15	1830	1809
	Part 16	1831	1813
	Part 17	1798	1814
	Part 18	1833	1816
	Part 19	1815	1791
	Part 20	1830	1812
	Part 21	1820	1799
옥누몽(서울대도서관본)	Part 1	1883	1874
	Part 2	1913	1899
	Part 3	1869	1866
	Part 4	1887	1866
소현성록(서울대도서관본)	Part 1	1833	1826
	Part 2	1830	1829
	Part 3	1845	1842
	Part 4	1846	1857
	Part 5	1833	1827

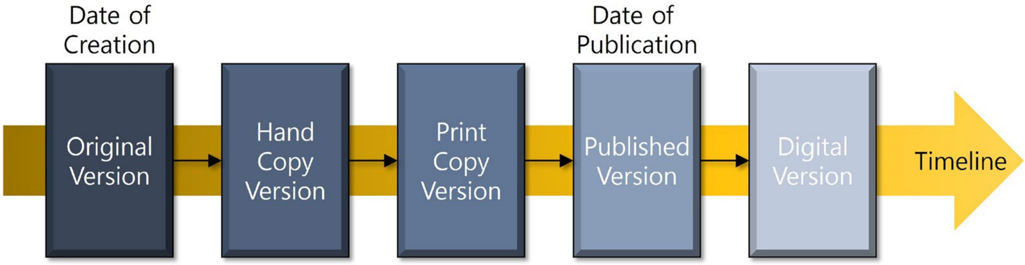

Fig. 6 The transition process from creation to digitization

grammar of Korean are not only limited to old Korean. For example, old Korean documents mainly use an ending (suffix, 어미) '하야', and it becomes '하여' in contemporary Korean. It is also obvious that the all Korean model contains more information, therefore shows higher accuracy. However, we have also tested the model considering all Koreans even with a low TF ratio, and the model shows higher test/validation loss, which means if the model considers characters with low TF, overfitting occurs while training.

If one desires to perform fine prediction within specific range, limiting the input data to the corresponding era would be available. However, as the number of databases is limited, additional documents with exact publicized/created date are believed to be helpful for increasing the accuracy of the model.

We note that the text pre-processing of our works is only based on preliminary syllable analysis. Recent techniques such as tokenization, word-embedding, multi-task learning, and Bidirectional Encoder Representations from Transformers (BERT) [31–34] have not been prepared for old Korean characters yet. If the text tokenization or embedding is available for old Korean, one could remove input nouns such as the name of character and places, etc., and perform an intensive study based on corpus. However, the simplicity of our method enables it to be applied to various languages. As our method only extracts the term frequency of each character or syllable, the study has the potential to be expanded to predict the created year of the document. The method based on the character is expected to be applicable to east Asian letters such as Chinese character, which has number of characters. Even for alphabet-based languages, data mining based on syllables would have the potential on estimating the created era or be useful for other data-based research.

Conclusion

In this paper, DNN is used to estimate the created era of old Korean documents. By processing the documents, raw, log, and binary TFs of Korean characters have been extracted and used as an input of the DNN model. Documents with a high proportion of Chinese character, and with a low proportion of Old Hangul has been excluded, and 496 documents have been the training, test, and validation set. In addition, Korean characters with low TF have been dismissed, 637 old Korean characters and 1218 contemporary Korean characters were considered to be the input of the model. Two DNN models were considered, one is the old Korean model, which only considers old Korean characters. The other is the all Korean model, which considers all Korean characters as input. Both models were trained and tested.

The accuracy of the test and validation set was high, and in particular, all Korean model showed an MAE of 13.77 years for test sets and 15.8 years for validation sets, which is at most a 3.24 % difference considering the total interest range of 487 years. The validity of the model was also verified by applying the model to documents where the approximate creation ages are estimated, where the predicted results of the model were in the expected range. Also, we have predicted different parts of identical novels, and they also showed similar predicted created years.

Predicting the created age of literary works using AI offers an indication to understand the literary trends and social conditions of the period by providing approximate information about the documents in which the created years are unknown. Furthermore, the predicted created year contributes to capturing the macroscopic flow of literary history. The data extraction method and AI model are expected to be applied not only to Hangul, but also to syllable-based text languages.

Abbreviations

AI: Artificial intelligence; NLP: Natural language processing; DNN: Deep neural network; TF: Term frequency; MLP: Multilayer perceptron; ReLU: Rectified linear unit; RMS: Root-mean-square; MSE: Mean-squared-error; MAE: Mean-absolute-error.

Acknowledgements

The two authors are married, however, not involved in a financial conflict of interest.

Author Contributions

IY designed the study, provided literary materials, and wrote about the significance of literature. HK performed data extraction and artificial intelligence application. All authors read and approved the final manuscript.

Funding

National Research Foundation of Korea (NRF) (2021R1F1A1052193); 2022 Hongik University Research Fund.

Availability of data and materials

The database of "21st century Sejong Project" is not owned by the authors, so raw data is unavailable to be provided from the authors. The database is owned by National Institute of Korean Language, Republic of Korea.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Korean Language and Literature, Seoul National University, Seoul, Korea. [2]Electrical and Electronic Convergence Department, Hongik University, Sejong, Korea.

Received: 21 July 2022 Accepted: 14 August 2022
Published online: 13 September 2022

References

1. Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Mach Learn. 1999;37(3):277–96.
2. Larochelle H, Bengio Y, Louradour J, Lamblin P. Exploring strategies for training deep neural networks. J Mach Learn Res. 2009;10(1):1–40.
3. Steinkraus D, Buck I, Simard P. Using GPUs for machine learning algorithms. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE; 2005; p. 1115–1120.
4. Schaller RR. Moore's law: past, present and future. IEEE Spectr. 1997;34(6):52–9.
5. Mack CA. Fifty years of Moore's law. IEEE Trans Semicond Manuf. 2011;24(2):202–7.
6. Lee H, Kwon H. Going deeper with contextual CNN for hyperspectral image classification. IEEE Trans Image Process. 2017;26(10):4843–55.
7. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98.
8. Kim H. Convolution Neural Network based Mode Decomposition for Degenerated Modes via Multiple Images from Polarizers. arXiv preprint arXiv:2207.03489. 2022.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. vol. 30. Curran Associates, Inc.; 2017.
10. Kamath U, Liu J, Whitaker J. Deep learning for NLP and speech recognition. vol. 84. Springer; 2019.
11. Pastor GC, Mitkov R, Afzal N, Pekar V. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In: Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers; 2008. p. 75–81.
12. Boroş T, Dumitrescu ŞD, Burtica R. NLP-Cube: End-to-end raw text processing with neural networks. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies; 2018. p. 171–179.
13. Dutoit T. High-quality text-to-speech synthesis: an overview. J Electr Electron Eng Aust. 1997;17(1):25–36.
14. Balyan R, McCarthy KS, McNamara DS. Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. In: Proceedings of the 10th International Conference on Educational Data Mining. ERIC; 2017.
15. Moreno-Jiménez LG, Torres-Moreno JM, et al. MegaLite: a new Spanish literature corpus for NLP tasks. In: Computing Conference; 2021.
16. Christou D, Tsoumakas G. Extracting semantic relationships in Greek literary texts. Sustainability. 2021;13(16):9391.
17. Kang WK, Kim B. Stylistics Consideration of <Sohyeonseongrok> series (소현성록 연작의 문체론적 고찰). Humanities Science Research (인문과학연구). 2018;59:29–46.
18. Yoo I, Kim H. Preliminary study on data analysis of Korean classical novels "Focused on Myeongjubowolbing and Yunhajeongsammunchwirok (한글장편소설의 데이터 분석에 대한 시론-＜명주보월빙＞,＜윤하정삼문취록＞을중심으로). The Society for Korean Language and Literary Research (문연구). 2022;50(2):175–200.
19. Li J, Song Y, Zhang H, Chen D, Shi S, Zhao D, et al. Generating classical Chinese poems via conditional variational autoencoder and adversarial training. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018. p. 3890–3900.
20. Fan H, Du W, Dahou A, Ewees AA, Yousri D, Elaziz MA, et al. Social media toxicity classification using deep learning: real-world application UK Brexit. Electronics. 2021;10(11):1332.
21. Wellek R, Warren A, et al. Theory of literature, vol. 15. Brace & World New York: Harcourt; 1956.
22. Carter R, McRae J. The Routledge history of literature in English: Britain and Ireland. Routledge; 2016.
23. Jung B. The status and characteristics of Classical Chinese and Vernacular Korean in Chosun period (조선시대한문과 한글의 위상과 성격에 대한 一考). Korean Culture (한국문화). 2009;48:3–20.
24. Paltoglou G, Thelwall M. A study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the 48th annual meeting of the association for computational linguistics; 2010. p. 1386–1395.
25. Jing LP, Huang HK, Shi HB. Improved feature selection approach TFIDF in text mining. In: Proceedings. International Conference on Machine Learning and Cybernetics. vol. 2. IEEE; 2002. p. 944–946.
26. Lee I, Ramsey SR. The Korean Language. Suny Press; 2001.
27. Moon SJ. A fundamental phonetic investigation of Korean monophthongs. Malsori. 2007;62:1–17.
28. Lee CH, Taft M. Subsyllabic structure reflected in letter confusability effects in Korean word recognition. Psychon Bull Rev. 2011;18(1):129–34.
29. Davis M, Collins L. Unicode. In: 1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings. IEEE; 1990. p. 499–504.
30. Kee MS. Translation of the Bible in Hangul. The Oxford Handbook of the Bible in Korea. 2022;p. 23.
31. Moon S, Okazaki N. Jamo pair encoding: Subcharacter representation-based extreme Korean vocabulary compression for efficient subword tokenization. In: Proceedings of the 12th Language Resources and Evaluation Conference; 2020. p. 3490–3497.
32. Jin K, Wi J, Kang K, Kim Y. Korean historical documents analysis with improved dynamic word embedding. Appl Sci. 2020;10(21):7939.
33. Aldjanabi W, Dahou A, Al-qaness MA, Elaziz MA, Helmi AM, Damaševičius R. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In: Informatics. vol. 8. MDPI; 2021. p. 69.
34. Hwang S, Kim D. BERT-based classification model for Korean documents (한국어 기술문서 분석을 위한 BERT 기반의 분류모델). J Soc e-Business Stud (한국전자거래학회지). 2020;25(1):203–214.