The Advancements and Progresses of Artificial Intelligence-Based Keyword Extraction Methods

Hexuan Deng

Jerudong International School, Bandar Seri Bagawan BE4719, Negara Brunei Darussalam
email: justin.deng@jis.edu.bn

Abstract. Keywords can provide general and significant information about documents; they allow readers to overview the text before deciding whether to read through or not. However, manually selecting keywords from texts faces problems such as time consumption, so efficient automatic keyword extraction methods are required. In this paper, the author discussed four different Artificial Intelligence (AI) models that were designed to complete the task of keyword extraction. Two of the models introduced used machine learning methods, and deep learning was adopted in the other two. The machine learning methods include decision tree and the Kea algorithm, both of them were supervised trained. The former uses the C4.5 algorithm to generate decision trees, and the latter uses the Naïve Bayes model to put words into classes. Deep learning models used Artificial Neural Networks (ANN) and the Long Short-Term Memory (LSTM). The ANN model consists of weighted layers that take four features of a word as inputs and returns one value that determines the class of the word. The LSTM model uses two networks, and the results from the networks are combined and passed into an attention layer to produce a vector that is to be used to classify the words. The models have problems with interpretability and distribution of data, which further works can be done on expert systems and domain adaptation to enhance the models from these problems. The paper summarizes the progress and potential improvements to the field of AI-based keyword extraction and can be used as reference for studying.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Keyword Extraction.

1 Introduction

Keywords are words or short phrases that are significant to a piece of information. They are often used by users to overview the contents of the information and determine whether they would like to use it or not. Accurate keywords may improve the efficiency and quality of information retrieval [1]. Correspondingly, faulty keywords may result in time wastage while searching for desired information. Thus, a precise keyword extraction method is crucial to information retrieval. Traditional methods of keyword extraction usually require manual input of keywords, which

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_61

happens to be time consuming and can be somehow inaccurate. In contrast, a relatively smaller amount of time is required for a trained artificial intelligence to produce accurate results, therefore it is a good method for keyword extraction.

Artificial intelligence and deep learning technologies have been developing rapidly in recent years, many different types of algorithms are developed for machines to learn and complete specific tasks. An example is the Random Forest algorithm [2], it functions by randomly choosing parts from a data set and letting different decision trees to handle these parts, then combine the results to create a final result. Another type of AI algorithm is neural network, which works by simulating human brains [3]. A neural network consists of weighted layers that act like neurons in the human body, where results from the previous layer are passed to the next layer and the result obtained from the last layer determines the final decision. These types of artificial intelligence are currently used in a variety of areas, for instance: chemistry, medicine, and especially in the field of information retrieval. A famous example of AI's application on information retrieval is ChatGPT, launched by OpenAI. In [4], researchers evaluated the ability of ChatGPT to retrieve information under a zero-shot setting, the results were satisfying and suggest the good potential of the model. Zhang et al. in [1] built a neural network model to extract keywords by modelling the contextual information of each given word; the relativity between the target word and the sentence before and after it is tested to determine if the word is identified as a keyword. A very popular approach to keyword extraction is the PageRank algorithm. PageRank is a successful algorithm in the information retrieval field, often used by browsers to return webpages that are relative to the searched questions; it is also often discussed when attempting to extract keywords from a piece of information [5-7]. This paper will provide a comprehensive conclusion to these new technologies, due to the significance of this field, and as it is developing quickly.

The remainder of this research is arranged into 3 main sections. First, the methodology of four AI models will be analyzed in section 2. Then, possible weaknesses and future obstacles of these works will be discussed in section 3. Finally, in section 4, the conclusion to this research will be drawn from what was discussed here.

2 Methodology

2.1 Machine Learning Methods

C4.5 Decision Tree. Turney P. D. proposed a method of using the C4.5 decision tree induction algorithm [8] for the extraction of keyphrases [9]. Supervised learning was used to train the algorithm to generate an accurate decision tree that can be used to decide the significance of a phrase in a piece of text. In order to create sets of training data, a list of phrases that consists of 1-3 words was made. The phrases were then stemmed using the Lovins stemmer and a featured vector was generated for each phrase. The features contained basic information about the phrase, and one feature (class) showed if the phrase is considered a keyphrase. The feature vectors were fed to C4.5 as training data sets to produce the decision tree.

The decision tree consists of root, decision, and leaf nodes; the root and decision nodes will test on the features provided for one phrase, and the tests will eventually lead to a leaf node that indicates the class of the phrase, hence deciding whether the phrase is a keyphrase or not. Soft-threshold decision trees were used by Turney in [8], this type of decision tree generates the probability for one phrase to be a keyphrase, so only the few phrases with the highest probability can be decided as the keyphrase. The bagging method was also used to reduce the variance of the outcome, the results from each tree were combined by taking averages to the probability; this method reduces the variance of the result.

The Kea Algorithm. Kea was put forward by Witten et al in [10], based on the Naïve Bayes algorithm [11]. The Naïve Bayes algorithm is often used to determine the likeliness of a given object to be in a class, by applying Bayes' rule to the sample data. The result from estimation can be used to classify the object, hence can be used to determine whether a word is a keyword.

The Kea algorithm consists of two main steps: identifying candidate phrases and calculating features. In the first step, the algorithm cleans the input text by splitting it into sentences, then it chooses candidate phrases by considering all the subsequences in the sentences. The candidates are further proceeded by stemming and case-folding for further calculations. Two features are calculated to classify the keywords. One of the features is a measure of the frequency of appearance of a phrase in the text compared to the rarity of the phrase. The other feature is the first occurrence, which indicates the position that the phrase first appears in the text.

Kea uses a supervised training method, similar to C4.5 decision tree in 2.1.1, which the classifications of the phrases are given when training to produce a model that is suitable for classifying phrases into keyphrases and non-keyphrases.

2.2 Deep Learning Methods

ANN. The neural network simulates human brains and consists of weighted layers that act like neurons. Layers in a neural network are split into three main layers: the input, hidden, and output layers. Units in input layer take are fed with attributes that are measured from the sample; weighted outputs are then calculated using these data and fed to the hidden layer for further calculation. There can be one or more hidden layers in a neural network, in each hidden layer the weighted outputs are calculated and passed on. At the final hidden layer, the results from the previous layer are fed to the output layer to generate a final outcom0065.

ANN are trained using back propagation. During the training process, the result from the network after each run is compared with the actual product, and the weights at the layers of the network are updated to minimize the difference between the result and the actual value. This process will be iterated until the weights start to converge; at this point the training is complete.

In the keyword extraction model proposed by Wang et al in [12], 4 features of a phrase are produced before the input to the neural network. The 4 features are: how frequently it appears in the sample, the inverted document frequency of the phrase, whether the phrase appears in the title of the sample text, and the paragraph distribution frequency of the phrase. After passing these features into the neural network model, a single output will be produced at the end. This output indicates the class of the phrase; the phrase is a keyword if the value is 1.

LSTM. Tang et al described their approach of using the LSTM model to perform keyword extraction in [13]. An LSTM network relies on the state of its cells, which are updated based on four gates. These four gates include the forget gate that is used to remove insignificant data received from the previous states; the input gate to determines the data that is to be updated; the input modulation gate to generate a vector to be added; the output gate determines the output from the cell. Tang et al in [13] used a modified version of LSTM, which is the BiLSTM. This model includes two networks, and their results are combined to give a final output.

An attention layer was used on top of the BiLSTM in [13], the purpose of this layer is to calculate the attention weight using a matrix. The attention weight reflects the relationship of the current word and the other words in the text, hence indicates the significance of the word. The output from the BiLSTM and the attention layer is used to calculate a context vector and this vector is used for final classification of the word.

3 Discussion

In [12], Wang et al stated a limitation to the neural network model, that the keyword extraction model is unable to obtain a word that has not appeared in the sample document, even if there is a strong relevance between the word and the text, examples to this include synonyms of the keywords. This may affect the accuracy of information retrieval: words that are synonymous with the keywords are searched, but the search results would not return all of the desired documents.

Other limitations include interpretability of the AI model. The interpretability of these models can be dissatisfactory, such that the reasons and results of an extraction of keyword cannot be explained. The model may show a good performance on the task. But without clear interpretations of its processes, it is difficult to find and refine possible weaknesses in the model. Thus, complex investigations may be required to study the model while trying to make updates, which is time consuming and can be costly.

Distribution of data is also a great limitation to artificial intelligence models. The models are trained in a certain environment with a limited data distribution, which means that the trained model may only be able to extract keywords from the distribution that it was trained with. For instance, a model trained with articles about computer science may show an exceedingly poor accuracy when it is used to extract keywords from a story book. Hence, the adaptability of the models is an area to be improved.

Expert systems can be used to improve the interpretability of a model. An expert system is an AI program that stores knowledge about a particular field and is used to simulate the behaviour of human experts. Expert systems are able to provide advice and explanations to users about how a result is obtained, thus, it can be used to enhance the interpretability of a model. The system can be applied to a keyword extraction model by providing it with the knowledge related to the text and then use it to generate explanations that help users to interpret the model.

In order to solve the problems on distributions, future works can be done on the area of transfer learning and domain adaptation. Transfer learning is a technique used specifically to solve the problem on the limitation in data distribution, it reuses existing models and knowledge to improve the performance of a new task [14]; relations between tasks are created to enhance the performance of the current task. Domain adaptation is a case in transfer learning, where only the domains of the data are changed, and the task required to perform remains constant [15, 16]. In the case discussed in this paper, the models only perform one task, which is to extract keywords, hence domain adaptation can be used to improve the models. Model applied with the method of domain adaptation may then be able to extract keywords accurately from samples that have different data distribution to the training data set.

4 Conclusion

In this paper, a review is made about the field of keyword extraction using Artificial Intelligence methods. Four different models using machine and deep learning methods were reviewed about: C4.5 Decision Tree and the Kea Algorithm use machine learning; the ANN and the LSTM model use deep learning methods. Limitations to the models were also discussed, including problems in the interpretability of the model and the data distribution in the training data sets. The problem of interpretability may impact on the maintenance and enhancement to the model, making it more difficult for improvements to be made. Limited training data results in inaccuracy of the keywords extracted when the models are used in different data distributions. Future works to be done may include using the expert system to enhance the interpretability of the models. Other works may be proposed by using the domain adaptation method on the models, in order to overcome the limitation on data distribution, and to allow for the models to perform precise extraction of data from a variety of data domains.

References

1. Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., Liu, T.: Keywords extraction with deep neural network model. Neurocomputing 383, 113–121 (2020).
2. Rigatti, S.J.: Random forest. Journal of Insurance Medicine 47(1), 31–39 (2017).

3. Dongare, A.D., Kharde, R.R., Kachare, A.D.: Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT) 2(1), 189–194 (2012).
4. Zhang, J., Chen, Y., Niu, N., Liu, C.: A preliminary evaluation of chatgpt in requirements information retrieval. In: Proceedings of the 2023 arXiv preprint arXiv:2304.12562 (2023).
5. Wang, J., Liu, J., Wang, C.: Keyword extraction based on pagerank. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 857–864. Springer, Berlin, Heidelberg (2007).
6. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2016).
7. Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., Hu, J.: Patent keyword extraction algorithm based on distributed representation for patent classification. Entropy 20(2), 104 (2018).
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier (2014).
9. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2, 303–336 (2000).
10. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical automatic keyphrase extraction. In: Proceedings of the Fourth ACM Conference on Digital Libraries, pp. 254–255 (1999).
11. Webb, G.I., Keogh, E., Miikkulainen, R.: Naïve Bayes. Encyclopedia of Machine Learning 15(1), 713–714 (2010).
12. Wang, J., Peng, H., Hu, J.S.: Automatic keyphrases extraction from document using neural network. In: Advances in Machine Learning and Cybernetics: 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, Revised Selected Papers, pp. 633–641. Springer Berlin Heidelberg (2005).
13. Tang, M., Gandhi, P., Kabir, M.A., Zou, C., Blakey, J., Luo, X.: Progress notes classification and keyword extraction using attention-based deep learning models with BERT. arXiv preprint arXiv:1910.05786 (2019).
14. Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., Azim, M.A.: Transfer learning: a friendly introduction. Journal of Big Data 9(1), 102 (2022).
15. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE, pp.877-894 (2020).
16. Qiu, Y., Hui, Y., Zhao, P., Wang, M., Guo, S., Dai, B., Dou, J., Bhattacharya, S., Yu, J.: The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process. Fuel 372, 132162 (2024 Sep 15).