A voice search engine for military symbols to enhance the drafting of operational plan documents on digital map

Dang Duc Thinh, Nguyen Chi Thanh, Le Van Diep,
Nguyen Duc Vuong, Luong Dinh Ha, Phung Nhu Hai*

Military Institute of Information Technology, AMST.
*Corresponding author: hainda59@gmail.com
Received 02 Mar. 2023; Revised 24 Apr. 2023; Accepted 10 May 2023; Published 25 May 2023.
DOI: https://doi.org/10.54939/1859-1043.j.mst.87.2023.40-49

ABSTRACT

The process of searching for information to serve the construction of operational plan documents on a digital map is still being done manually and needs to be automated in order to improve efficiency. Speech recognition and natural language processing technologies, commonly used in chatbots, virtual assistants, voice commands, and voice search, could be promising tools to overcome this problem. This paper proposes a framework for deploying a voice search engine that uses Whisper, a deep learning-based automatic speech recognition model, and combines TF-IDF, N-gram, and Truncated SVD as feature extraction approaches to search for text ground truth in a dictionary of military symbols using Cosine similarity. Despite the small size of a custom dataset, the experiments show promising results, achieving an accuracy of 82.00%. Our achievement surpasses that of several traditional statistical methods and classification models.

Keywords: Voice search; Feature extraction; Cosine similarity; Military symbols; Digital map.

1. INTRODUCTION

The software to support the operation of operational plan documents on digital maps developed by the Military Information Technology Institute - Academy of Military Science and Technology, such as TMHQ, T3BD provides a set of supporting solutions for drafting, presenting, and reporting digital maps-based operational plan documents [1]. This practical toolkit has been applied in practice and has received positive user feedback.

The current process of manually searching for information to build operational plan documents is time-consuming and relies on a single individual. To improve efficiency, reduce working time, and increase flexibility, some tasks in the process should be automated and made available through multiple working channels.

In recent years, speech recognition and natural language processing technologies have exploded and become indispensable to people's daily lives. Automatic Speech Recognition (ASR) technology allows the computer to automatically convert speech into text, helping users not need to type and still be able to enter data [2]. Natural language processing (NLP) is a technology related to the processing, analysis, and understanding of natural human language by computers [3]. These two technologies are often combined to create intelligent applications such as chatbots, virtual assistants, voice commands, and automatic voice searches.

The voice search function can be easily found on virtual assistants such as Alexa, Siri, and Google Assistant [4]. Users can use voice to search for information or perform tasks on devices equipped with these virtual assistants. Then, the virtual assistants will query online search services such as Google, Bing, YouTube, or Wikipedia to find information related to the user's request.

In the military field, the voice search feature is often integrated with control and information systems on weapons and combat vehicles to provide information to commanders and units directly in combat.

Unfortunately, these technologies are confidential and difficult to access by the militaries of other countries. Automation control systems developed by domestic defense units mainly stop at solving professional problems and have yet to apply automatic voice search technology in operational support.

Therefore, in this paper, we propose to build an automatic search engine for military symbols by voice. The main contributions of the research are twofold:

- Constructed own audio dataset with many people recording and converted it into text data using an ASR model for testing the proposed method.
- Proposed a framework that combines character-level natural language processing methods with TF-IDF, N-grams, and truncated SVD to vectorize text data and calculate Cosine similarity to classify and complete the search content while resisting input interference.

The rest of the paper is structured as follows: Section 2 reviews related works on voice control applications. Section 3 and section 4 describe the proposed method and experiments with them. The results and discussions are presented in section 5, and finally, conclusions and future works are provided in section 6.

2. RELATED WORKS

Voice control applications are divided into two research directions, Voice to Commands and Voice to Text.

Recent research in the field of Voice to Commands has produced quite impressive results. For example, research by Majumdar et al. has proposed a neural network architecture called "MatchboxNet" composed of blocks of 1D time-channel separable convolution blocks, batch-normalization, ReLU, and dropout layers to recognize spoken commands with the best accuracy on the Google Speech Commands dataset [5]. Kim et al. introduced a deep learning method called "broadcasted residual learning" to recognize keywords on smart devices [6]. To achieve high accuracy with a small model size and low computational load, the method employs 1D convolution calculations with residual connections. Additionally, the method incorporates broadcasted residual connections to reduce computational load further. Oleg Rybakov et al. published research focusing on keyword detection models in mobile phone streaming and non-streaming modes [7].

However, the unique feature of Voice to Commands technology is that it focuses on classifying commands or keywords from voice, not converting voice into text. The disadvantage of this technology is the limitation on the number of statements that can be recognized, the content of the statement is often shortened without considering the semantics, especially since it does not work well in an environment with noise, many sound sources, and volume changes. Therefore, it is unsuitable for the problem we are trying to solve.

Speech to Text research direction can overcome the above characteristics. Some research works on the application of Speech to text have been published, typically:

The work of Egozi et al. on concept-based information retrieval, generated using explicit semantic analysis rather than keyword-based search, addresses inaccuracies or omissions in the search results that may arise from different keywords being used to describe the same concept [8].

In 2018, Ravanelli et al. proposed a simplified model of Gated Recurrent Units (GRUs) called Light GRU (Li-GRU) for ASR [9]. The Li-GRU model reduces the training time per epoch by more than 30% compared to a standard GRU and increases the accuracy of continuous speech recognition across various tasks, different noise, and input conditions.

Vietnamese research group Anh D. T. et al. proposed an ASR model using transformer architecture and won third place in the Vietnamese Speech and Language Processing competition in 2021 with a word error rate (WER) of 8.83% on the private-test dataset [10].

The above works have proved that Speech to text is a feasible method to solve the problem of complexity and a large amount of input audio content by converting audio to text and processing semantics. However, no research has been conducted on speech-to-text processing for automatic search purposes in the military field. This is what motivated us to conduct this study.

3. PROPOSED METHODS

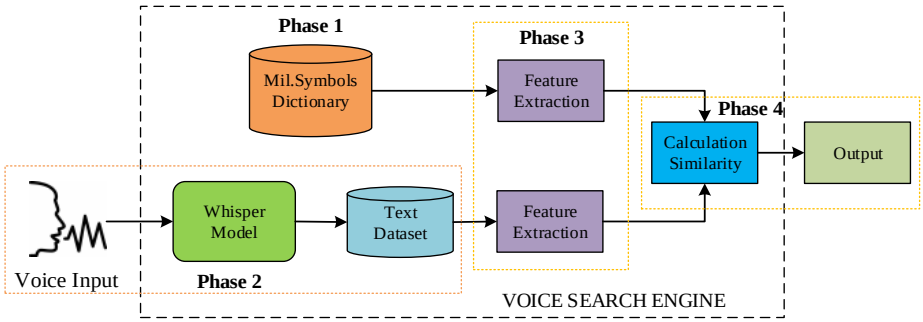

Figure 1. *The proposed framework for searching military symbols from speech input.*

In order to automatically search for military symbols by voice, the paper proposes a framework consisting of phases:

- *Phase 1*: Building a dictionary of military symbols;
- *Phase 2*: Converting voice input to text dataset;
- *Phase 3*: Extracting appropriate features for text dataset and dictionary of military symbols;
- *Phase 4*: Calculating the similarity between the input search vectors and the vectors from the dictionary using Cosine Similarity, selecting the highest value in the similarity score vector, and returning the text ground truth.

The overall framework procedure is illustrated in figure 1.

3.1. Building a dictionary of military symbols

The military symbol dictionary has been built from the 100 most commonly used military symbols in missions related to sea and island defense drills, where the software

for drafting operational plan documents on digital maps is being effectively used and promoted. Each symbol has a unique label (text ground truth) according to the operational standards issued by the General Staff of the Vietnamese People's Army [11]. Several of the military symbols in the dictionary are provided in table 1, column 'Text ground truth'.

3.2. Converting voice input to text dataset

With a large number of military symbols to search and the need to perform well in noisy conditions, it is necessary to convert the input voice data into text for processing in the following steps instead of using Voice to command technology.

This research uses OpenAI's Whisper model for converting voice into text [12]. Whisper is a versatile automatic speech recognition model created by OpenAI, trained on a large and diverse dataset (over 680,000 hours) to help perform speech recognition and language translation tasks. Whisper works well in noisy conditions and can convert audio into numeric characters instead of words like other ASR models. Finally, we chose Whisper because it is an open-source model where OpenAI shares all the source code, allowing the creation of a separate application that can run independently on a workstation or LAN, while other models are only provided as an API.

For building the dataset, we used 50 symbols from the built-in 100-symbol dictionary to give to 6 different people to read and record. This ensures the generalizability of the problem in real-world conditions where searching through a very large set of symbols is required, but the number of symbols required for each type of operational document is much smaller. Each symbol was recorded to a file, corresponding to the use case when searching for a symbol for the operation. The data from 6 voice recorders were fed through the Whisper model to obtain a Dataset consisting of 300 text samples. This dataset is labeled with the text ground truth in the dictionary. We used 250 data samples corresponding to 5 people for training and 50 samples of the remaining person for testing.

Table 1. *Some examples of the output of Whisper and their text ground truths. Bold indicates incorrect words with the text ground truth.*

Text ground truth	**Output of Whisper**	**CER (%)**
ký hiệu tàu đánh cá nước ngoài *foreign fishing vessel symbol*	ký hiệu tàu đánh cá nước ngoài *foreign fishing vessel symbol*	00.00
ký hiệu tàu đánh cá quốc doanh *state-owned fishing vessel symbol*	**kỷ** hiệu tàu đánh cá quốc doanh *state-owned fishing boats century-sign*	03.30
ký hiệu tàu vận tải dân sự *civil transport ship symbol*	ký **ước** tàu vận tải dân **xe** *civilian-vehicle transport ship contract*	24.00
ký hiệu tàu tên lửa loại nhỏ *small missile ship symbol*	ký **họ tấu** tên **nữa** loại nhỏ *they play name again small type sign*	28.00
ký hiệu tăng hạng nhẹ *light tank symbol*	ký **ước tâm hận nhạc** *a wish to hate music sign*	57.90

In table 1, examples of Whisper's output with corresponding text ground truth are presented, highlighting the presence of noise. To assess this noise, we calculate the Character Error Rate (CER) using the following formula:

$$CER = \frac{(S + D + I)}{N} \cdot 100\% \quad (1)$$

where, S is the number of incorrect characters (substitution); D is the number of missing characters (deletion); I is the number of redundant characters (insertion); N is the total number of characters in the correct text.

Based on the statistics in table 1, it can be observed that the text ground truth exhibits common characteristics such as short sentences, simple sentence structure, repeated words, and distinguishing information placed on one side of the sentence. The output of Whisper encounters some level of noise, ranging from 0 to 60%. There are sentences with heavy noise, which are almost identical to the text ground truths in terms of intonation. The origin of these noises can be attributed to several factors, such as human pronunciation errors, errors during recording, and inaccuracies in the Whisper model while processing Vietnamese. Vietnamese is a monosyllabic language, where a word consists of one or multiple syllables. Each syllable changes its meaning when the tone changes.

3.3. Extracting appropriate features for text dataset and dictionary of military symbols

To extract features of text data obtained from the Whisper model and text ground truth in the dictionary of military symbols, we perform the following steps:

- *Step 1*: Converting text to vector using Char level TF-IDF combined with N-gram;
- *Step 2*: Dimensionality reduction of the vector created with Truncated SVD.

Based on the frequency of occurrence of words in a document and their inverse frequency in the entire dataset, Term Frequency-Inverse Document Frequency (TF-IDF) is used to evaluate the importance of words or symbol characters in a text [13]. Meanwhile, the method of dividing a text into paragraphs (N-grams) can help capture the structure and correlation between words in a sentence [14]. Truncated Singular Value Decomposition (SVD) is used to reduce the dimensionality of feature vectors, making storage and computation more efficient [15]. This is particularly useful for long and sparse text vectors, which often have large dimensions but few nonzero values.

When combining character-level and N-gram TF-IDFs, we obtain a feature vector of the text, which includes information about the frequency of occurrence of each character along with information about the relationship between those characters. This will solve the problem of word noise in terms of syllables. Then, further combining with SVD helps to improve the accuracy of text classification and search models, especially when dealing with short and specific documents such as military symbols.

3.4. Calculating the similarity between the input search vectors and the vectors from the dictionary using Cosine Similarity

To recover search information that is affected by noise in the text, an effective method is to compute the similarity between the input vector and all vectors from the military symbol dictionary. Two calculation methods can be considered: Euclidean distance and Cosine similarity [16].

When calculating in a multidimensional space, the distance between two vectors may not fully indicate their degree of similarity. Cosine similarity provides a more relative measure that examines the similarity between two vectors by measuring the angle between them. Additionally, cosine similarity effectively deals with the removal of the

impact of the vector's length. Using Euclidean distance, the distance between two vectors is greater if they have different lengths. Meanwhile, cosine similarity preserves the similarity between two vectors regardless of their length. For the above reasons, we choose Cosine to calculate the similarity.

Cosine similarity is calculated as follows:

$$cos(u, v) = \frac{\langle u, v \rangle}{\| u \| \cdot \| v \|} \quad (2)$$

where, u and v are two feature vectors of two sentences that must be compared for similarity; $\| u \|$ and $\| v \|$ represent the lengths of vector u and v, respectively (norm according to $L2$ or Euclidean norm). The resulting value ranges from -1 to 1, where -1 indicates two opposite text sentences, and 1 indicates two completely similar sentences.

After calculating the similarity, a text sentence from the voice will be predicted to correspond to a dictionary sentence with the greatest similarity.

4. EXPERIMENTS

4.1. Experimental Settings

Our experiments were conducted using Python 3.9, along with the PyTorch and Scikit-Learn libraries. The experiments were performed on a computer equipped with an Intel Xeon CPU (3.40 GHz) and an NVIDIA Quadro RTX 4000 GPU.

4.2. Evaluation Metrics

To evaluate the performance of the proposed model, we used one metric as follows:

$$Accuracy = \frac{C}{T} \cdot 100\% \quad (3)$$

where, C is the number of correctly predicted text sentences with the label, T is the total number of predicted sentences.

4.3. Competing feature extraction approaches and classification models

To demonstrate the effectiveness of the proposed method, we used several approaches to vectorize the input text, such as Bag of Words, Word-level TF-IDF with and without a combination of N-gram and SVD, and Word2vec. Furthermore, we trained two popular classification models, KNN (K-Nearest Neighbors) and SVM, to compare with our proposed method.

5. RESULTS AND DISCUSSIONS

The comparisons of the proposed framework with the competing methods are presented in table 2 and figure 2. From the results in table 2, it can be seen that the proposed method achieves the highest accuracy by 82.00% with a competitive average search time per symbol of 1.179 milliseconds.

Table 2. *The comparison results of the proposed method with other methods.*

No	Group	Method	Accuracy (%)	AST (ms)
1	Feature Extraction Approaches	W2V + SVD	68,00	0.985
2		W-BOW	72.00	1.084
3		W-BOW + SVD	76.00	0.959

4		C-BOW	74.00	1.100
5		C-BOW + SVD	76.00	0.965
6		W-TFIDF	68.00	1.084
7		W-TFIDF + SVD	74.00	0.966
8		W-TFIDF + N-gram + SVD	76.00	1.008
9		C-TFIDF	76.00	0.977
10		C-TFIDF + SVD	78.00	0.997
11		C-TFIDF + N-gram + SVD	**82.00**	1.179
12	Classification Models	C-TFIDF + N-gram + K-NN	78.00	**0.134**
13		C-TFIDF + N-gram + Multi-class SVM	72.00	1.515

Abbreviation: W2V – Word2vec; W-BOW – Word level Bag Of Words; C-BOW – Char level Bag Of Words; W-TFIDF – Word level TF-IDF; C-TFIDF – Char level TF-IDF; SVD – Truncated Singular Value Decomposition; AST – Average Search Time per symbol; ms – milliseconds.

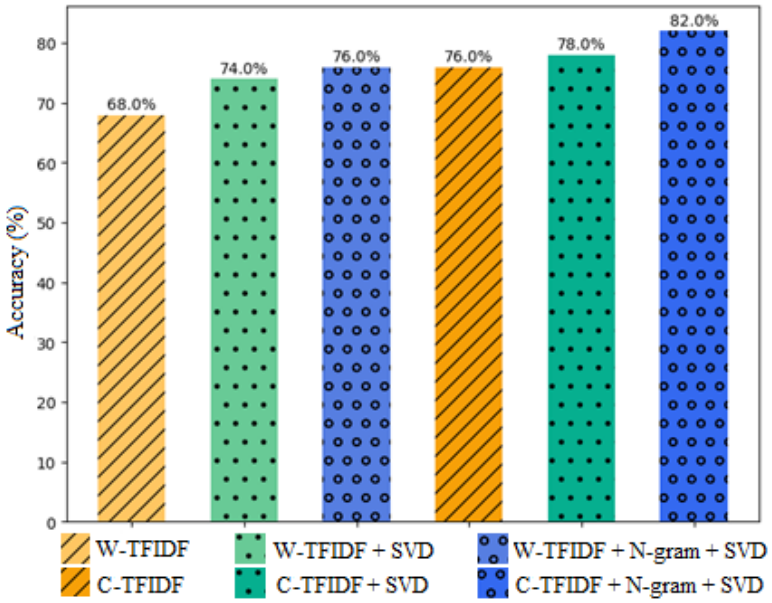

Figure 2. *Accuracy increases with the combination of TF-IDF, N-gram, and SVD.*

The graph in figure 2 shows that the method's accuracy increases when combining the three techniques: TF-IDF, N-gram, and SVD, and increases when switching from word level to char level. The results prove the hypothesis that using a combination of traditional statistical methods to extract features for short texts with frequent word repetitions and taxonomic meanings at the end of sentences, where syllabic noise is prevalent, is perfectly reasonable.

The SVD hyperparameter *n_components* determines the number of principal components retained after applying the Truncated SVD method to analyze the original vector. These principal components can be used to represent the original data in a new space with a lower dimension.

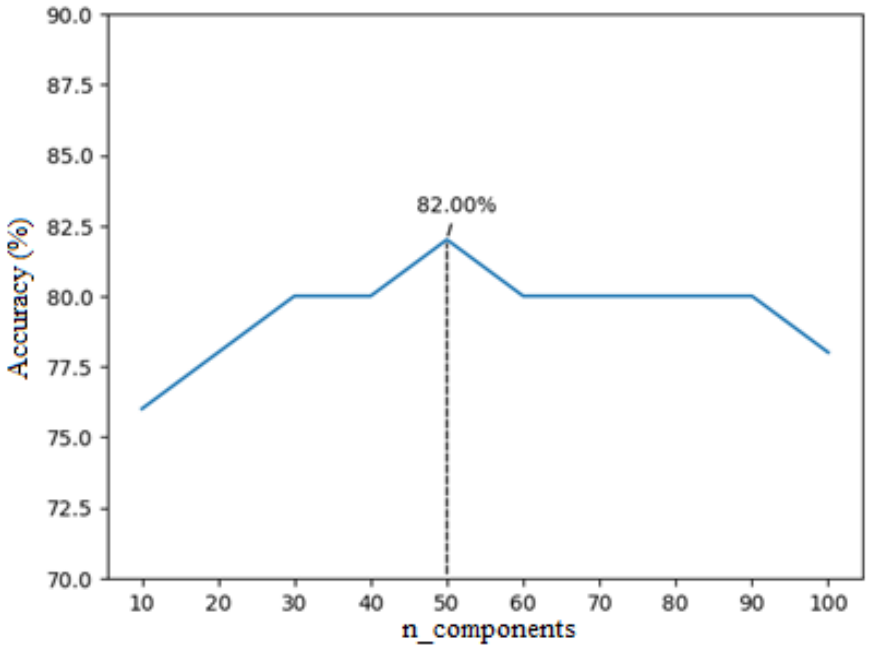

Figure 3. *Accuracy graph of the SVD hyperparameter (n_components).*

From figure 3, the method results are highest when the SVD hyperparameter *n_components* is set to 50. When *n_components* is less than 50, the amount of information lost in the feature vectors causes a decrease in classification results. When *n_components* is greater than 50, the length of the feature vector increases and the number of zero values also increases, leading to a decrease in the accuracy of the results.

6. CONCLUSIONS AND FUTURE WORKS

This paper proposes building an automatic search engine for military symbols using voice input, utilizing the Whisper model for voice-to-text conversion, and using feature extraction approaches, such as Char-level TF-IDF, N-gram, Truncated SVD, and Cosine similarity calculation methods for searching text ground truth in a dictionary of military symbols. We compared our framework with several traditional statistical methods and equivalent machine learning models, and the results confirmed the effectiveness of the proposed method.

In future works, we plan to test our framework with a larger dataset, compare the model in more detail, and explore more comprehensive methods for improving the proposed framework. Based on that, we will integrate this tool into the software for drafting operational plan documents on digital maps in the military symbology operational module. Additionally, we will work on improving the process of automatic symbol writing, drawing, editing, and automatically adding information to combat objects on digital maps.

Acknowledgement: *This research is part of a scientific research project for young staff in 2023 of the Academy of Military Science and Technology: "Researching and building a voice recognition tool applying artificial intelligence to automate information search in the construction of battle plan documents on sea and island warfare on digital maps".*

REFERENCES

[1]. Nguyen Duc Dinh, Hoang Van Toan, *"System Design Documentation of T3BD System"*, (2020).

[2]. Arthur Brown, *"How Does Voice Recognition Work?"*, (2021) [Online]. Available: https://www.makeuseof.com/how-does-voice-recognition-work.

[3]. Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, Freeson Kaniwa, Kutlwano Sedimo, *"Natural Language Processing: A Review"*, International Journal of Research in Engineering and Applied Sciences, vol. 6, is. 3, (2016).

[4]. Raul Mercado, *"Siri vs. Alexa vs. Google Assistant: Which Is Smarter at Answering Questions?"*, (2021) [Online]. Available: https://www.makeuseof.com/siri-vs-alexa-vs-google-smarter-answering-questions.

[5]. Somshubra Majumdar, Boris Ginsburg, *"MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition"*, Audio and Speech Processing (eess.AS), (2020), doi: https://doi.org/10.21437/Interspeech.2020-1058.

[6]. Byeonggeun Kim, Simyung Chang, Jinkyu Lee, Dooyong Sung, *"Broadcasted Residual Learning for Efficient Keyword Spotting"*, Sound (cs.SD), (2021), doi: https://doi.org/10.48550/arXiv.2106.04140.

[7]. Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirko Visontai, Stella Laurenzo, *"Streaming keyword spotting on mobile devices"*, Audio and Speech Processing (eess.AS), (2020), doi: https://doi.org/10.21437/Interspeech.2020-1003.

[8]. Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich, *"Concept-Based Information Retrieval Using Explicit Semantic Analysis"*, ACM Transactions on Information Systems, vol. 29, is. 2, pp. 1–34, (2011), doi: https://doi.org/10.1145/1961209.1961211.

[9]. Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, Yoshua Bengio, *"Light Gated Recurrent Units for Speech Recognition"*, IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, is. 2, pp. 92-102, (2018), doi: 10.1109/TETCI.2017.2762739.

[10]. Viet Duong Trinh Anh, Sam Dang Van, Tuan Do Van, Vi Ngo Van Trong, *"Vietnamese Automatic Speech Recognition with Transformer"*, EasyChair Preprint, no. 7147, (2021).

[11]. General Staff, *"Military Symbols"*, People's Army Publishing House, (2021).

[12]. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, *"Robust Speech Recognition via Large-Scale Weak Supervision"*, Audio and Speech Processing (eess.AS), (2022).

[13]. Shahzad Qaiser, Ramsha Ali, *"Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents"*, International Journal of Computer Applications, vol. 181, no. 1, (2018), doi: 10.5120/ijca2018917395.

[14]. William Cavnar, John M. Trenkle, *"N-Gram-Based Text Categorization"*, Environmental Research Institute of Michigan, (2001).

[15]. Serge Shishkin, Arkadi Shalaginov, Shaunak D. Bopardikar, *"Fast approximate truncated SVD"*, Numerical Linear Algebra with Applications, vol. 26, no. 1, (2019), doi: 10.1002/nla.2246.

[16]. Dani Gunawan, C A Sembiring, Mohammad Andri Budiman, *"The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents"*, Journal of Physics Conference Series, vol. 978, no. 1, (2018), doi: 10.1088/1742-6596/978/1/012120.

TÓM TẮT

Một công cụ tìm kiếm ký hiệu quân sự bằng giọng nói phục vụ xây dựng văn kiện tác chiến trên nền bản đồ số

Việc tìm kiếm thông tin phục vụ xây dựng văn kiện tác chiến trên nền bản đồ số vẫn đang được thực hiện thủ công, cần được tự động hóa để tăng hiệu quả sử dụng. Công nghệ nhận dạng giọng nói và xử lý ngôn ngữ tự nhiên, thường được sử dụng trong chatbot, trợ lý ảo, ra lệnh bằng giọng nói và tìm kiếm bằng giọng nói, có thể giúp tự động hóa một số tác vụ. Bài báo này đề xuất xây dựng một công cụ tìm kiếm tự động các ký hiệu quân sự bằng giọng nói, sử dụng mô hình Whisper để chuyển đổi giọng nói thành văn bản, các phương pháp xử lý ngôn ngữ tự nhiên như TF-IDF, N-gram, Truncated SVD và phương pháp tính độ tương đồng Cosine được dùng để hoàn thiện thông tin tìm kiếm bằng từ điển kí hiệu quân sự. Chúng tôi đã so sánh phương pháp đề xuất với một số phương pháp thống kê truyền thống và mô hình máy học tương đương. Mặc dù bộ dữ liệu âm thanh mà chúng tôi nghiên cứu thu thập rất hạn chế, các thử nghiệm cho thấy kết quả tốt với độ chính xác là 82,00%. Kết quả này cao hơn những phương pháp trích xuất đặc trưng truyền thống và các mô hình phân loại, từ đó, khẳng định tính hiệu quả của phương pháp đề xuất.

Từ khoá: Tìm kiếm bằng giọng nói; Trích xuất đặc trưng; Độ tương đồng Cosin; Ký hiệu quân sự; Bản đồ số.