Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

An Applied Study of English Sentence Fast Retrieval Algorithm Based on Choquet Expectation

Tingting Liu[1,] ✉

[1] *Basic Teaching Department, Henan Polytechnic, Zhengzhou, Henan, 450046, China*

ABSTRACT

In this paper, the basic structure of fuzzy integral-based multi-classifier fusion model is used as a reference to construct Choquet integral vectors, measure the similarity of English sentences, and construct a fast retrieval algorithm for English sentences based on Choquet expectation. Determine the algorithm threshold and compare the running time of similar retrieval algorithms. Deploy the algorithm into the English sentence retrieval model for dataset training and comparison experiments. Verify the model robustness and determine the chosen K value for the model. Further use the test set to compare the retrieval effectiveness of the model with the traditional semantic retrieval model. The algorithm threshold is set to 6 to improve English sentence recall. The running time consumption of the algorithm is 0.827s and 1.941s, which is lower than the other three similar retrieval algorithms. In the dataset comparison experiments, the algorithmic model of this paper scores better than the comparison model in all 5 evaluation metrics. The model has the best robustness when k takes the value of 15. The model check accuracy and check completeness are higher than the semantic retrieval model LM by nearly 8 percentage points. The fast retrieval algorithm for English sentences based on Choquet expectation can improve sentence retrieval timeliness and retrieval accuracy, and reduce retrieval energy consumption.

Keywords: fuzzy integration, multi-classifier fusion, integration vector, fast retrieval algorithm, Choquet expectation

1. Introduction

Since the 21st century, human beings have stepped into the era of informationization and intelligence. As the most important bearer of information, natural language, how to efficiently solve the language barrier between different languages has become an important topic that human society cannot ignore. Machine translation is an effective way to solve this problem by realizing automatic switching between multiple languages with the help of computers. Machine translation is a technology that uses

✉ Corresponding author.
E-mail address: 22036@hnzj.edu.cn (T. Liu).
Received 25 February 2024; Revised 10 June 2024; Accepted 05 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-158

computers to translate one natural language into another target language. Due to the complexity of natural language, machine translation has become one of the ten most difficult problems in today's scientific and technological world [1-2]. Although the quality of machine translation translations is not perfect at present, machine translation is useful, and it has been used in many aspects such as language teaching, communication, information retrieval, cross-border e-commerce cooperation and so on since its birth [3-6].

Today, the era of big data and cloud computing has arrived, communication between different languages is more and more common, and the barrier between languages is becoming more and more prominent, and machine translation, as the most important means of overcoming language barriers, will become more and more obvious in modern society. The development of machine translation has formed numerous branches technically speaking, including direct, transformational (rule-based) [7], pivot (based on intermediate languages) [8], corpus-based (statistical, instance-based) [9-10] and hybrid strategies (multi-engine) [11]. Corpus-based approaches are gaining acceptance due to the fact that with the help of automatic computer processing of realistic example sentences in the corpus, it is possible to extend limited rules to an infinite extent, thus avoiding the need to write translation rules manually [12-13]. In the corpus-based approach, the example-based machine translation system directly takes the translation of similar examples in the corpus as a template, and generates the final translation after necessary corrections, which makes full use of the resources of the corpus and thus avoids the need for obscure syntactic analysis of the sentences [14-15]. Considering that there may be more than one sentence instance similar to the sentence to be translated and the size of the corpus, the retrieval of the most similar sentence instances becomes one of the key techniques in the instance-based machine translation system [16]. The translation quality of an instance-based approach is affected by whether the parallel corpus contains similar instances or not. In practice, the size of the corpus should be as large as possible [17]. However, the larger the size, the longer the retrieval time of similar instances, the higher the retrieval accuracy requirement, and the real-time and accuracy cannot be guaranteed [18]. Thus, the study of similar sentence instance retrieval algorithms in the corpus is a guarantee to improve the overall performance of the instance machine translation system, and the retrieval of similar sentences essentially belongs to the category of information retrieval. In the field of information retrieval, traditional retrieval techniques often rely on the degree of keyword matching, and the more matches there are, the greater the degree of similarity between two entities is considered to be. In contrast, in today's retrieval techniques, it is more advocated to measure the degree of difference between two entities with the help of inter-sentence similarity [19-21]. Sentence similarity can be divided hierarchically into three levels: syntactic, semantic, and pragmatic [22]. Syntactic similarity is to syntactically analyze two sentences, build up two syntactic trees, and calculate the similarity between two sentences syntactically, however, its defect of considering only syntax but not semantics dooms it to be used only as a reference data. Semantic similarity refers to the degree of overlap between two sentences in the surface meaning, and its function is higher than the syntactic similarity but lower than the pragmatic similarity [23]. Semantic similarity is the degree of similarity between two sentences in practical applications, which is also the highest level pursued by machine translation systems, however, it is extremely difficult to realize, and is only in the theoretical stage [24]. In practical applications, the semantic similarity between sentences is more often used. However, whether it is any level of syntax, semantics, or pragmatics, its sentence retrieval still needs to pursue speed while ensuring quality.

In this paper, the specific principles of fuzzy measures and fuzzy integrals are elaborated, and the specific architecture of fuzzy integral-based multi-classifier fusion model is analyzed. The fast retrieval algorithm of English sentences based on the fuzzy integral of words to calculate the actual similarity of sentences is designed through two steps of constructing Choquet integral vector and measuring the similarity of English sentences. The threshold of the algorithm is determined by the assistance of accuracy and recall. The algorithm is experimentally compared with edit distance-based, same-word-

based and vector-based TF-IDF methods alone to analyze the advantages of the algorithm in reducing the running time. Through the deployment of the algorithm, the application of the fast English sentence retrieval algorithm based on Choquet's expectation in real English sentence retrieval models is investigated. Compare the performance difference with similar retrieval models by combining the dataset and evaluation metrics. Validate the model robustness and further compare the differences with semantic retrieval models.

2. Overview

Word Frequency-Inverse Document Frequency (TF-IDF) is a representative search for words or phrases with high frequency of occurrence in one text and low frequency of occurrence in other texts to retrieve the text, so it has been used by a number of scholars for sentence retrieval. Literature [25] extracted stems and lexical reduction in sentences by TF-IDF method using deactivation removal and language modeling techniques and performed sentence retrieval of varying lengths taking into account both stems and lexemes. Literature [26] accelerates similar case retrieval in case-based machine translation process with simhash algorithm for fast English sentence retrieval. Similarly, literature [27] study mentioned that in considering the rules and corpus of the machine translation system, sentence similarity retrieval with simhash algorithm, the process maintains the precision and accuracy while the retrieval performance of similar words is improved. Literature [28] uses statistical language model to extract featured vocabulary in the corpus, after calculating the frequency of vocabulary combined with ontology to reduce the understanding difference between machine and natural language, in order to create a fast retrieval model for English sentences, and form a new algorithm after adding machine learning technology, which has improved the performance of the algorithm in the smaller size of the corpus. Literature [29] uses open linguistic resources as an aid to extract features in a semantic kernel and aggregates these features using weights generated by supervised machine learning techniques, and an intelligent text retrieval engine for sentence similarity is built in these contexts. Literature [30] utilizes an edit distance algorithm for distributed retrieval of similar sentences in English, where retrieval is no longer a single node in a large-scale sample repository, making retrieval efficiency soar. Literature [31] fused semantic word answer generation algorithm and cuckoo search optimization algorithm with sentence similarity retrieval based on semantic level to improve the accuracy of sentence retrieval. The above studies contributed to sentence similarity retrieval and were mostly conducted based on word frequency, but the speed, precision, and accuracy of retrieval varied due to different corpus sizes. For uncertain corpus size, sentence length and other fuzzy problems when the fast retrieval of sentences puts forward higher requirements. While Choquet expectation has an advantage in dealing with uncertainty and fuzzy new problems, it considers the interrelationships and importance of each numerical element through a nonlinear weighting function [32].

3. Principles of modeling fuzzy measures and fuzzy integral correlations

In this section, based on the multi-classifier fusion architecture, fuzzy measures and fuzzy integration methods are combined to construct a fuzzy integration-based multi-classifier fusion model. The base classifiers are utilized to learn the samples and output non-negative real vectors. The output vectors of all base classifiers form a decision matrix. Calculate the fuzzy integral of a column in the decision matrix with respect to the fuzzy measure, and get the vector value of the samples in that column. The superscript of the largest vector value is the classification result for that sample.

3.1. Fuzzy Measurement

Before delving into fuzzy integrals, it is important to understand the concept of fuzzy measures. A fuzzy measure is a measure of the “importance” or “influence” of a set, which, unlike traditional probability measures, is able to capture more nuanced and complex information, especially when dealing with

uncertainty and interactions between classifiers. For a better understanding, we can imagine fuzzy measures as an evaluation system for teamwork, where the contribution (influence) of each team member (classifier) to the success of a project depends not only on their individual competencies, but also on how they cooperate with other team members.

For example, consider a simple situation where a project team consists of three members, and we use a fuzzy measure to evaluate how much they contribute to the success of the project. Here, the fuzzy measure not only reflects each member's individual contribution, but also expresses the fact that a certain two members can have more impact working together than when working alone, which is difficult to capture with traditional measures.

Define fuzzy measure: μ is said to be a fuzzy measure on (A,B) if and only if μ satisfies the following conditions:

1) (Triviality) $\mu(\phi)=0$ if $\phi \in B$;

2) (Monotonicity) $\forall X \in B$, $Y \in B$ and $\mu(X) \le \mu(Y)$ if $X \subseteq Y$;

3) (lower continuity) if $X_n \subset B(n=1,2,\cdots,\infty)$, $X_1 \subset X_2 \subset \cdots \subset X_n \subset \cdots$, and $\bigcup_{n=1}^{\infty} X_n \in \mathrm{B}$, then

$$\lim_n \mu(X_n) = \mu\left(\bigcup_{n=1}^{\infty} X_n\right) \tag{1}$$

4) (Upper continuity) $\forall \{X_n\} \subset B(n=1,2,\cdots,\infty)$, $X_1 \supset X_2 \supset \cdots \supset X_n \supset \cdots$, and $\bigcap_{n=1}^{\infty} X_n \in \mathrm{B}$, then

$$\lim_n \mu(X_n) = \mu\left(\bigcap_{n=1}^{\infty} X_n\right) \tag{2}$$

A fuzzy measure μ is said to be a regular fuzzy measure when $\mu(A)>0$; μ is said to be a lower semicontinuous fuzzy measure when μ satisfies only the conditions of ordinariness, monotonicity and lower continuity; and μ is said to be an upper semicontinuous fuzzy measure when μ satisfies only the conditions of ordinariness, monotonicity and upper continuity.

3.2. Fuzzy (Choquet) integrals

Choquet integral is a flexible integration method that can effectively deal with non-additive measures and is suitable for decision-making and classification problems in a variety of scenarios. Since the Choquet integral takes into account various influencing factors and also emphasizes the role of interrelatedness and mutual constraints among indicators on the evaluation results, this paper adopts the Choquet integral as the fusion operator of the model.

Defining the Choquet integral: let f be a nonnegative real-valued measurable function defined on a finite set X and μ be a fuzzy measure defined on X, then f the Choquet integral with respect to the fuzzy measure μ is defined $(c)\int f d\mu$ as:

$$(c)\int f d\mu = \int_0^{\infty} \mu(\{x \mid f(x) > \gamma\}) dl \tag{3}$$

where the integral on the right-hand side of the equation is the integral of a function of γ with respect to the Leberger measure.

When $X=\{x_1,x_2,\cdots,x_n\}$ is a finite set, $f: X \to [0,1]$, the formula for the Choquet integral becomes accordingly:

$$(c)\int f d\mu = \sum_{i=1}^{n} \left(f(x_i) - f(x_{i-1})\right)\mu(A_i) \tag{4}$$

where $A_i = \{x_i, x_{i+1}, \cdots, x_n\}$, $f(x_0) = 0$.

Without loss of generality, assume $f(x_1) \le f(x_2) \le \cdots \le f(x_n) \le 1$, (otherwise, the element in the rearrangement is $\{x_1^*, x_{i+1}^*, \cdots, x_n^*\}$ such that the corresponding function value $f(x_i^*)$ satisfies $f(x_1^*) \le f(x_2^*) \le \cdots \le f(x_n*) \le 1$).

3.3. Basic structure of fuzzy integral-based multi-classifier fusion models

Fig. 1 shows the basic structure of the fuzzy integral-based multi-classifier fusion model. Assume that a multi-classifier fusion system dealing with a c-class classification problem has L base classifiers, the set of classifiers is denoted as $D = \{D_1, \cdots, D_L\}$, and the set of samples has n samples, denoted as $X = \{x_1, \cdots, x_n\}$.

Learning to classify the sample set using L base classifier respectively. In a fuzzy integral-based multi-classifier fusion system, the most basic requirement is that the outputs of the base classifiers are non-negative real vectors. For a c-class classification problem, the output vector of the ird classifier D_i for sample x_k is $p_{ik} = \left[p_{ik}^1, \cdots, p_{ik}^j, \cdots, p_{ik}^c\right](i = 1, \cdots, L; k = 1, \cdots, n)$, where $p_{ik}^j \in [0,1](j = 1, \cdots, c)$, denotes the likelihood that classifier D_i will classify sample x_k into category j. For sample x_k, the outputs of the L base classifiers form a decision matrix DP_k:

$$DP_k = \begin{bmatrix} p_{1k}^1 & \cdots & p_{1k}^j & \cdots & p_{1k}^c \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{Lk}^1 & \cdots & p_{Lk}^j & \cdots & p_{Lk}^c \end{bmatrix} \tag{5}$$

A fuzzy measure μ^j can be added to the category $C_j (j = 1, \cdots, c)$ defined L on the power set of the set $D = \{D_1, \cdots, D_L\}$ of classifiers.

The j th column $p_k^j = \left[p_{1k}^j, \cdots, p_{Lk}^j\right]^T$ in the decision matrix DP_k is considered as a function f^j on the set D, and the fuzzy integral $P_k^j = \int f^j d\mu^j$ on the fuzzy measure μ^j is computed f^j, P_k^j i.e., the likelihood of categorizing the sample x_k into a class C_j obtained by fusing the predictions of the L base classifiers by the fuzzy integral. For the sample x_k, the prediction result obtained by the fusion model of the classifiers based on fuzzy integration is a c dimensional vector $P_k = \left[P_k^1, \cdots, P_k^c\right]$ and the superscript of the value of the largest element in the vector is the final classification result for the sample x_k.

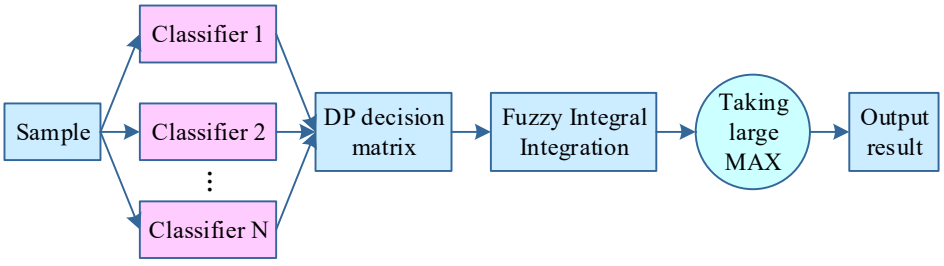

Fig. 1. Basic structure of multi-classifier fusion model based on Choquet integral

4. Algorithm design and validation

This section combines the basic structure of the Choquet integral-based multi-classifier fusion model in the previous section to design a fast retrieval algorithm for English sentences based on Choquet expectation. The threshold of the algorithm is determined using the corresponding English sentence corpus. The algorithm is compared with other sentence retrieval algorithms in terms of running time to verify the performance level of the algorithm in terms of reducing the English sentence retrieval time.

4.1. English Sentence Similarity Measure Based on Choquet Integral Vector

4.1.1. Constructing the Choquet Integral Vector. Step1: There is a sentence S, which is divided into words by the English lexical system to obtain the word $A_1, A_2, \cdots, A_m$, then the vector formed by the integral word A_i contained in the sentence is called the integral word vector of the sentence S, denoted as $S = \{A_1, A_2, \cdots, A_m\}$.

Step2: In the integral vector of sentence S, the integral word A_{i-1} before the integral word A_i is called the A_i pre-integral word, and the integral word A_{i+1} after the integral word A_i is called the A_i post-integral word.

Step3: The length of the integral vector in sentence S is $Len(S) = m$, initialize each integral word A_i weight $w_i = 1/m$ and all the weight values form a vector of weight values denoted as $SW = \{w_1, w_2, \cdots, w_m\}$.

Step4: Given two sentences $S1$, $S2$ in the integral word vector $S1 = \{A_1, A_2, \cdots, A_m\}$, $S2 = \{B_1, B_2, \cdots, B_n\}$, $S1$ all the integral words that exist in $S2$ at the same time constitute a vector called $S1$ based on the existence of $S2$ vector, denoted as $E = \{A_1, A_2, \cdots, A_p\}$.

Step5: The weight values corresponding to the integral words in the existence vector E form $S1$ the existence value vector based on $S2$, denoted as $SE = \{v_1, v_2, \cdots, v_p\}$.

Step6: Calculate the similarity of $S1$ and $S2$ using the above given integral word vectors.

4.1.2. English sentence similarity measures. Let $S1$ and $S2$ be two sentences, $S1$ contains the word $A_1, A_2, \cdots, A_m$ and $S2$ contains the word $B_1, B_2, \cdots, B_n$, i.e.:

$$S1 = \{A_1, A_2, \cdots, A_m\}, S2 = \{B_1, B_2, \cdots, B_n\} \tag{6}$$

Assuming that the $S1$ vector is shorter in length than $S2$, i.e., $m < n$, the initial vector of weight values for $S1$ is computed:

$$SW1 = \left\{\frac{1}{m}, \frac{1}{m}, \cdots, \frac{1}{m}\right\} \tag{7}$$

For each integrator A_i in $S1$, if A_i is present in $S2$, see if its preintegrator in $S1$ and $S2$ is the same, and if it is, expand the value of the corresponding weight in SW1 by a factor of σ. See if its postintegrator in $S1$ and $S2$ is the same, and if it is, also expand the value of the corresponding weight by a factor of σ. If A_i does not exist in $S2$, the corresponding weights remain unchanged.

Then compute $S1$ based on the existence vector of $S2$:

$$E = \{A_1, A_2, \cdots, A_p\} \tag{8}$$

For each integral word in E, the corresponding weight values from the weight vector $SW1$ form the existence value vector SE:

$$SE = \{v_1, v_2, \cdots, v_p\} \tag{9}$$

Finally, the similarity of $S1$ and $S2$ based on the integral vectors is calculated according to Eq. (9).

$$VectorSim(S1, S2) = \frac{\sum_{i=1}^{p} v_i}{\sum_{i=1}^{m} w_i} \times \frac{m}{n} \tag{10}$$

4.2. Threshold determination

The threshold value in the fast retrieval algorithm for English sentences based on Choquet's expectation is itself an empirical value, and its determination needs to be based on the specific context of use. The size of the corpus used for the experiments in this paper is 9950 parallel English-Chinese sentence pairs, and another 32 English sentences in which similar instances can be found in the corpus constitute the test set. The algorithm threshold is determined on this basis.

The threshold is an empirical value and there is no specific formula for its calculation, but it can be determined with the aid of accuracy and recall in the field of information retrieval.

Accuracy rate is the ratio of the number of relevant texts retrieved to the total number of texts retrieved, which measures the rate of checking accuracy. The formula is: accuracy rate = number of correct texts extracted / total number of texts extracted.

Recall rate is the ratio of the number of relevant texts retrieved to the number of all relevant texts in the library, which measures the check accuracy rate. The formula is: Recall rate=number of correct texts extracted/total number of all relevant texts in the library.

In this paper, all the English sentences in the test set are tested and the accuracy and recall of each English sentence under different thresholds are calculated, and then the average accuracy and average recall are calculated for each threshold. Figure 2 shows the threshold-accuracy-recall curve. For the retrieval results, people always hope that the higher the accuracy and recall at the same time, the better, but in fact the two are usually contradictory. From Fig. 2, we can see that the intersection point of the accuracy and recall curves is located on the right side of threshold 5. In order to consider the two indexes of accuracy and recall at the same time, the value should be set to 5. However, considering that the fast retrieval algorithm for English sentences based on Choquet's expectation is used to narrow down the range of possible similar instances, and does not directly produce the final result; at the same time, in order to prevent similar example sentences from being filtered due to problems such as synonyms, the threshold should be enlarged appropriately to improve the recall. Therefore, in this paper, the algorithm threshold is set to 6 instead of 5.

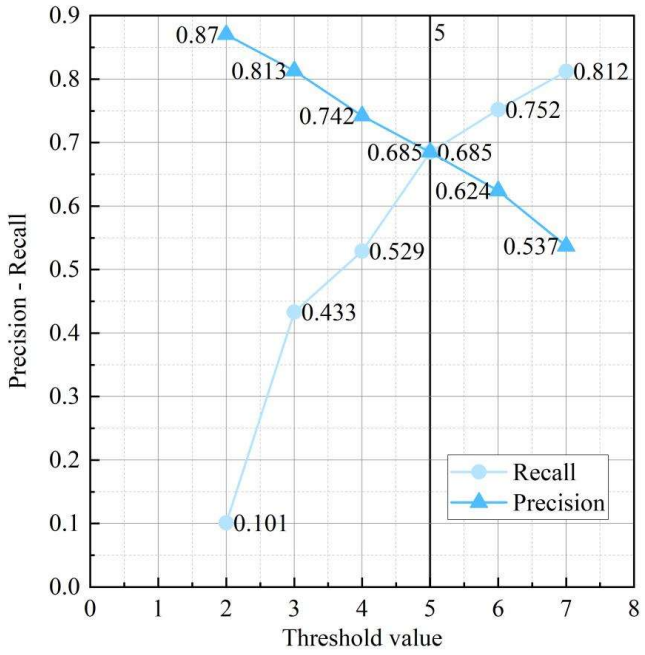

Fig. 2. Curve of threshold, accuracy and recall rate

4.3. Algorithm runtime comparison

In order to validate the fast English sentence retrieval algorithm based on Choquet's expectation designed in this paper, it is experimentally compared with three other commonly used retrieval algorithms - edit-distance based, same-vocabulary based, and vector-based TF-IDF method alone. The experiments are supported by the corpus used in this paper, and all the other three algorithms are reimplemented and compared in terms of runtime results. Thirty-two English sentences in the test set are tested with the last four algorithms, and the running time of each sentence is recorded to calculate the average elapsed time of the same algorithm.

Figure 3 shows the comparison results formed by the 4 algorithms when none of them deal with synonyms and all of them deal with synonyms. As can be seen from Fig. 3, the time consumption of the separate TF-IDF algorithm reaches 70.92s and 97.29s due to the need to construct a high-dimensional vector for each instance, which is much higher than that of the other three algorithms; the methods based on the editing distance and based on the same vocabulary do not have a big difference in terms of the running time; whereas the algorithm of the present paper has a time consumption of 0.827s and 1.941s, which is better than the other three algorithms.

The time consumption of all four algorithms increases after the introduction of synonym processing. In the method based on edit distance and the same vocabulary, only the synonym query process is added, while in the TF-IDF method, the calculation of lexical similarity is added separately, so the time consumption of the TF-IDF method is much higher than the previous two algorithms. The algorithm in this paper determines sentence similarity by first calculating word similarity through Choquet integral vector, which narrows the range of existence of similar instances and reduces the number of times of synonym query and word similarity calculation, so it reduces the impact of the synonym processing stage on the overall performance, while retaining the advantages of high precision and accuracy of the calculation results.

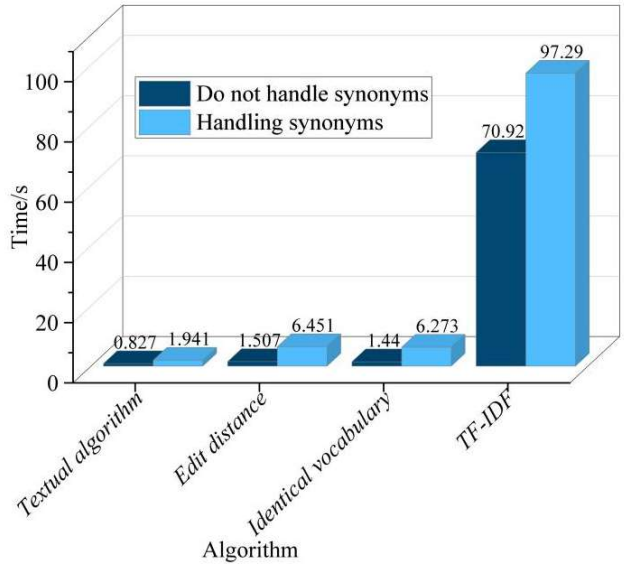

Fig. 3. Comparison of running time of the four algorithms

5. Algorithmic applications and analysis

In order to study the practical application value of the English sentence fast retrieval algorithm based on Choquet expectation designed in this paper, it is deployed in the English sentence fast retrieval model to be trained and analyzed. Comparative experiments are conducted on different retrieval models using datasets to analyze the results of this paper's algorithmic model on each evaluation index. Visualize the results of different values of k in dynamic k-max pooling to verify the robustness of the algorithmic model in this paper. Further design the corpus and test set to verify the advantages of this paper's algorithmic model over traditional semantic retrieval models.

5.1. Comparative Experiments on Data Sets

This paper uses the WebAP dataset. The dataset is constructed based on documents from TREC GOV2 as well as retrieval. Each English sentence in the dataset has been tagged accordingly, and each tagging result is mapped into a [4,3,2,1] score, which is very suitable for comparison experiments of retrieval models. In terms of evaluation metrics, this paper chooses the evaluation metrics commonly used in the field of information retrieval: NDCG@10, P@10, MRR, AP, Bpref. A total of four models in three categories were selected for comparison experiments: the best model for traditional information retrieval, the most recent model in the Q&A task category and the most recent model on the WebAP dataset. The four models specifically include: the language model LM, the convolutional neural network model CNN, the multi-view deep text matching model MV-LSTM, and the learning sequencing method L2R+CSFeatures. Among them, the learned ranking method L2R+CSFeatures achieved the best results on the WebAP dataset, and its retrieval method is the best among the four models.

Table 1 shows the experimental results for the WebAP dataset. This paper's model outperforms the results of other models on all evaluation metrics. Specifically, on the Top-1 evaluation metrics MRR, AP, and Bpref, this paper's model improves 28.56%, 12.98%, and 32.27%, respectively, over the current best model L2R+CSFeatures; while on the Top-10 evaluation metrics NDCG@10 and P@10, this paper's model improves 37.48% and 13.63%, respectively. Meanwhile, from Table 1, we can see that the effect of the models (L2R+CSFeature and the model in this paper) that use contextual sentence information is much higher than that of the models that do not use contextual sentence information, which indicates that for non-factual retrieval, the information provided by individual sentences is often not enough to give a complete answer, and better results can be given with the help of contextual

sentence information. At the same time, we also note that LM works better than the other two neural network-based baseline methods without using contextual sentence information, which may be due to the fact that the retrieval in the WebAP dataset contains rich keyword information, such as the search for "describe history oil industry", in which the exact matching information is far more effective than the approximate matching, and the ability to distinguish between exact and approximate matching, LM is more capable than CNN and MV-LSTM, so better results are achieved.

Table 1. Experimental results of WebAP dataset

Model	NDCG@10	P@10	MRR	AP	Bpref
LM	0.1341	0.1452	0.3396	0.1553	0.1642
CNN	0.0597	0.0647	0.1255	0.0748	0.0898
MV-LSTM	0.0778	0.1051	0.2223	0.1152	0.1079
L2R+CSFeatures	0.1865	0.2025	0.4513	0.2126	0.2166
Textual model	0.2564	0.2301	0.5802	0.2402	0.2865

5.2. Robustness Verification

In order to verify the robustness of the model in this paper, we visualize and analyze the results for the values of k taken in dynamic k-max pooling. Figure 4 shows the comparison of the results of different k values in this paper's model. From Fig. 4, it can be seen that the choice of different k values does not have a great impact on the English sentence retrieval results. When the value of k is equal to 105, the results are slightly decreased, mainly because the average length of retrieval in the dataset is 7, while the average sentence length is 16. Therefore, too large a value of k for most of the retrievals will introduce additional noisy information, which will affect the performance of the experiment. Meanwhile, we observe that the best results are achieved when k takes the value of 15.

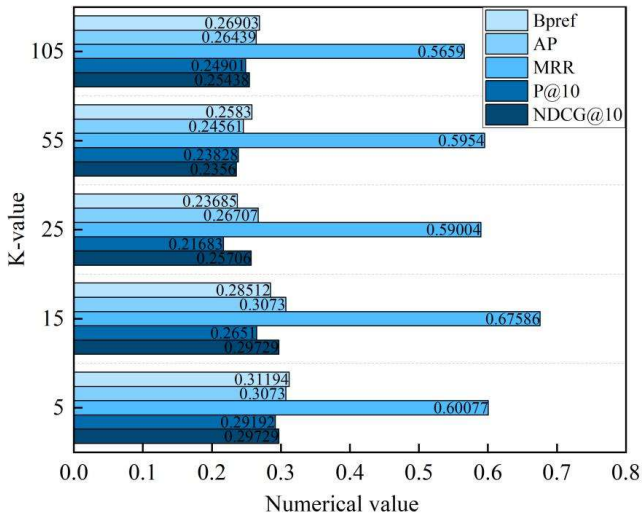

Fig. 4. Comparison of results of different K values in the model in this paper

5.3. Model Testing and Analysis of Results

Based on the k-value taken as 15, in order to further compare the advantages and disadvantages of the English sentence retrieval results of the traditional semantic retrieval model LM and this paper's model, 140 English sentences were selected from the corpus used in Chapter 3 to form a new test set, and the two models were tested again. Table 2 shows the comparison between the traditional semantic retrieval model LM and this paper's fast English sentence retrieval model based on Choquet's expectation. Analyzing Table 2, it can be seen that out of 140 test sentences, this paper's model

correctly matches 120 sentences, which is more than the 110 of the semantic retrieval model LM. The check accuracy and check completeness of this paper's model are 85.71% and 96% respectively, which are nearly 8 percentage points higher than the semantic retrieval model LM. It can be seen that the actual retrieval effect of calculating the similarity degree of contextual sentences through the fuzzy integral of words is better than the semantic retrieval method of exact matching.

Table 2. Statistics about the results of Word Sense LM and Textual model

Method	LM	Textual model
Number of test sentences	140	140
Contains the correct number of matching sentences	110	120
The system does not give the correct number of matching sentences	15	5
There are no matching sentences in the corpus	15	15
Accuracy rate P	78.57%	85.71%
Recall rate R	88%	96%
MMR	1.108	1.153

6. Conclusion

In this paper, we design an English sentence fast retrieval algorithm based on Choquet expectation, study its retrieval performance advantage of calculating the actual similarity of English sentences through the fuzzy integral of words, and verify the application value of the algorithm. With the aid of accuracy and recall determination, the algorithm is able to obtain the optimal English sentence retrieval effect when the threshold is taken as 6, avoiding the influence of synonyms on sentence similarity. Using the corpus as a support, comparing the difference in running time between this retrieval algorithm and the other three retrieval algorithms, it is found that the algorithm consumes 0.827s of time when it does not deal with synonyms, and 1.941s of time when it does. The algorithm outperforms the other three retrieval algorithms in terms of time performance, regardless of whether or not synonyms are processed. It is verified that the designed algorithm can effectively reduce the time required for English sentence retrieval and provide fast response to users.

The algorithm is deployed into the English sentence fast retrieval model and the dataset is utilized for training and comparison experiments. In the results of the five evaluation metrics, NDCG@10, P@10, MRR, AP, and Bpref, the sentence retrieval performance of the model is improved by 37.48%, 13.63%, 28.56%, 12.98%, and 32.27%, respectively, compared with the current best model, L2R+CSFeatures. It shows that the method based on word fuzzy integration to calculate contextual sentence similarity is significantly better than retrieval methods based on neural networks and exact matching. Visualizing the obvious results for different values of k, it is found that the model has the best robustness when k is taken as 15. Comparing this model with the traditional semantic retrieval model LM using the new test set, the check accuracy and check completeness of this paper's model are 85.71% and 96%, respectively, which are higher than that of LM by nearly 8 percentage points.

Sentence-level retrieval is of great importance in retrieval scenarios with more explicit information needs such as automatic translation and mobile applications. The fast English sentence retrieval algorithm based on Choquet expectation designed in this paper can significantly improve the timeliness and retrieval accuracy of English sentence retrieval, and the effect is more obvious when the corpus size is larger. The algorithm is applied to automatic question and answer, document abstraction, machine translation and other fields, which can assist the relevant personnel to quickly process English sentences and improve work efficiency.

References

[1] Rivera-Trigueros, I., Olvera-Lobo, M. D., & Gutiérrez-Artacho, J. (2021). Overview of machine translation development. In Encyclopedia of Information Science and Technology, Fifth Edition (pp. 874-886). IGI Global.

[2] Specia, L., & Shah, K. (2018). Machine translation quality estimation: Applications and future perspectives. Translation quality assessment: from principles to practice, 201-235.

[3] Hellmich, E. A. (2021). Machine translation in foreign language writing: Student use to guide pedagogical practice. Alsic. Apprentissage Des Langues et Systèmes d'Information et de Communication, 24(1).

[4] Khan, A., & Mirza, S. (2024). Global Voices: The Role of Machine Translation in Multilingual Communication. Eastern European Journal for Multidisciplinary Research, 1(1), 7-10.

[5] Shubham, M. (2024). Breaking Language Barriers: Advancements in Machine Translation for Enhanced Cross-Lingual Information Retrieval. J. Electrical Systems, 20(9s), 2860-2875.

[6] Li, Y., & Luo, B. (2024). Cross-Border E-commerce Product Introduction: Needs and Characteristics for Localized Machine Translation. Journal of Business and Management Studies, 6(3), 125-133.

[7] Khanna, T., Washington, J. N., Tyers, F. M., Bayatlı, S., Swanson, D. G., Pirinen, T. A., ... & Alos i Font, H. (2021). Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. Machine Translation, 35(4), 475-502.

[8] Le Thuyen, P. T., & Hung, V. T. (2015). Results Comparison of machine translation by Direct translation and by Through intermediate language. International Journal, 3(5).

[9] Ismailov, A. S., Shamsiyeva, G., & Abdurakhmonova, N. (2021). Statistical machine translation proposal for Uzbek to English. Science and Education, 2(12), 212-219.

[10] Wong, B. T. M. (2023). Example-based machine translation. Routledge Encyclopedia of Translation Technology, 145-159.

[11] Banik, D., Ekbal, A., Bhattacharyya, P., & Bhattacharyya, S. (2019). Assembling translations from multi-engine machine translation outputs. Applied Soft Computing, 78, 230-239.

[12] Luo, J., & Li, D. (2022). Universals in machine translation? A corpus-based study of Chinese-English translations by WeChat Translate. International Journal of Corpus Linguistics, 27(1), 31-58.

[13] Kouremenos, D., Ntalianis, K., & Kollias, S. (2018). A novel rule based machine translation scheme from Greek to Greek Sign Language: Production of different types of large corpora and Language Models evaluation. Computer Speech & Language, 51, 110-135.

[14] Tehseen, I., Tahir, G. R., Shakeel, K., & Ali, M. (2018, May). Corpus based machine translation for scientific text. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 196-206). Cham: Springer International Publishing.

[15] Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. Machine Translation, 32(3), 195-215.

[16] Rai, S., Belwal, R. C., & Gupta, A. (2023). Is the corpus ready for machine translation? A case study with Python to pseudo-code corpus. Arabian Journal for Science and Engineering, 48(2), 1845-1858.

[17] Navarro, S. (2024). Investigating Machine Translation Quality Across Genres: A Corpus-Based Analysis of Phraseology. In The Routledge Handbook of Corpus Translation Studies (pp. 600-619). Routledge.

[18] Jassem, K., & Dwojak, T. (2019). Statistical versus neural machine translation–a case study for a medium size domain-specific bilingual corpus. Poznan Studies in Contemporary Linguistics, 55(2), 491-515.

[19] Boban, I., Doko, A., & Gotovac, S. (2020). Improving sentence retrieval using sequence similarity. Applied Sciences, 10(12), 4316.

[20] Kadilierakis, G., Fafalios, P., Papadakos, P., & Tzitzikas, Y. (2020). Keyword search over RDF using document-centric information retrieval systems. In The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17 (pp. 121-137). Springer International Publishing.

[21] Hambarde, K. A., & Proenca, H. (2023). Information retrieval: recent advances and beyond. IEEE Access.

[22] Sun, X., Meng, Y., Ao, X., Wu, F., Zhang, T., Li, J., & Fan, C. (2022). Sentence similarity based on contexts. Transactions of the Association for Computational Linguistics, 10, 573-588.

[23] Sravanthi, P., & Srinivasu, B. (2017). Semantic similarity between sentences. International Research Journal of Engineering and Technology (IRJET), 4(1), 156-161.

[24] Chauhan, S., Kumar, R., Saxena, S., Kaur, A., & Daniel, P. (2024). Semsyn: Semantic-syntactic similarity based automatic machine translation evaluation metric. IETE journal of Research, 70(4), 3823-3834.

[25] Boban, I., Doko, A., & Gotovac, S. (2020). Sentence retrieval using stemming and lemmatization with different length of the queries. Advances in Science, Technology and Engineering Systems, 5(3), 349-354.

[26] Ouyang, J. (2020, July). Research on the fast retrieval algorithm of english sentences based on simhash. In 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 997-1000). IEEE.

[27] Lai, C. (2022). Fast retrieval algorithm of English sentences based on artificial intelligence machine translation. In 2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City: Volume 1 (pp. 1057-1065). Springer Singapore.

[28] Wang, H. (2023). A novel algorithm for the construction of fast English sentence retrieval model using a combination of ontology and advanced machine learning techniques. Soft Computing, 27(23), 18129-18146.

[29] Amir, S., Tanasescu, A., & Zighed, D. A. (2017). Sentence similarity based on semantic kernels for intelligent text retrieval. Journal of Intelligent Information Systems, 48, 675-689.

[30] Li, Y., Yan, Y., & Ye, J. (2019). Distributed Retrieval of English Similar Sentences Based on the Edit Distance. Journal of Computations & Modelling, 9(1), 55-64.

[31] Kanagarajan, K., & Arumugam, S. (2019). Intelligent sentence retrieval using semantic word based answer generation algorithm with cuckoo search optimization. Cluster Computing, 22(Suppl 3), 7003-7013.

[32] Chen, J., & Chen, Z. (2020). A new proof for the generalized law of large numbers under Choquet expectation. Journal of Inequalities and Applications, 2020, 1-17.