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

A New Strategy of Artificial Intelligence-Based CARS Model Applied to Comparative Analysis of Medical Paper Abstracts in Chinese and English

Peijun Liu[1,✉]

[1] *Department of General Education, Foreign Language Teaching and Research Section, West Anhui Health Vocational College, Lu'an, Anhui, 237000, China*

ABSTRACT
Artificial Intelligence (AI) is increasingly used in medical research, especially in the analysis and interpretation of medical data. In this study, based on the traditional CARS model, we built a framework for thesis abstract language step research by categorizing fuzzy steps into optional steps and adding appropriate key steps to the language steps. With the help of artificial intelligence technology, an extraction model of key elements of abstracts incorporating the attention mechanism is constructed, aiming at screening the elemental utterances in abstracts. Finally, by collecting data from medical related papers in CNKI, Web of Science and other databases, the CARS modeling strategy based on artificial intelligence is implemented in the comparative analysis of medical paper abstracts in English and Chinese. Through the comparative analysis, it is found that the number of sentences in English abstracts is concentrated in 6-7 sentences, while the number of sentences in Chinese abstracts is scattered in 2-8 sentences. The percentage of the use of Chinese sentences on English abstract writing is the highest, with an average percentage of 45.24%. The frequency of the first 20 words of fuzzy restrictive phrases in English abstracts was significantly higher than that in Chinese abstracts. The organization of Chinese and English abstracts was mostly in the structure of "introduction-method-results-discussion", which accounted for 54% and 71%, respectively. In addition, the frequency of steps indicating gaps in the research area is higher in English than Chinese abstracts.

Keywords: Artificial intelligence, CARS model, Attention mechanism, Element extraction model, Medical abstracts

1. Introduction

As the professionalism and innovation of dissertation writing are increasing, more and more dissertation writers, in addition to the innovation and professionalism of the research topic itself, improve the professionalism of writing in English and Chinese in the field of research, to ensure the quality of writing, and to prepare for the publication of dissertations in international journals.The

✉ Corresponding author.
E-mail address: princecom2012@163.com (P. Liu).
Received 05 March 2024; Revised 10 May 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-372

CARS model, which translates as “Creating a Research Space”, was proposed by Swales. The CARS model, proposed by Swales, summarizes the writing of a textual part of a dissertation into three moves, and a number of steps to achieve these moves [1]. Steps, i.e. actions and steps, are rhetorical structures or rhetorical patterns that convey a communicative function and guide the writer to write and the reader to understand the text, and are units of meaning with a clear purpose and function, consisting of different linguistic strategies [2]. The names of these three linguistic steps are Establishing the research area (M1), Establishing the research gap (M2), Occupying the research gap or Introducing the study (M3).M1 has three steps, i.e., indicating the importance of the research area, outlining the central theme, and summarizing the previous studies. M2 has two steps, i.e., pointing out the deficiencies of the existing studies or expanding and supplementing the existing ones, and providing the rationale for filling in the gaps. M3 has three steps, i.e., introducing the research area and the central theme, and providing the rationale for filling in the gaps. M3 also has three steps, i.e., introducing the purpose of the study or introducing the study, presenting the main findings, and indicating the structure of the article [3-4].

The abstract is an important part of the dissertation. It is a short text that provides a summary of the literature without comments and additional explanations, briefly and accurately describes the important contents of the literature, has independence and self-explanatory, and has the same amount of the main information as the literature, that is, without reading the whole text, you can get the necessary information, and it is economical to concentrate the essence of the full-text content [5]. After a scientific research paper is published, its abstract can be entered into the paper version or electronic version of the abstract search journals, which becomes an important basis for searching the original text and guides the reader to read the original text. The most common structure of a dissertation abstract is a four-step process, i.e., problem formulation, methodology, results, and conclusions [6].Swales found that the structure of a dissertation abstract is consistent with the structure of a dissertation, i.e., it is also structured as a four-step process of introduction-methodology-results-discussion. Early genre analysis of academic papers began to penetrate from the introduction and genre to the abstract, introduction, results, discussion and other parts of academic papers [7]. Nowadays, most of the papers have both English and Chinese abstracts, so that speakers of different languages can understand the main content of the article as soon as possible, and provide convenience for people to carry out literature searches and preliminary categorization. As the main method of genre analysis and genre research, speech step analysis is an important tool for analyzing dissertation abstracts.

Most researchers' use of CARS model focuses on the introduction part, for example, literature [8] constructs the introduction of the dissertation through CARS model, and at the same time analyzes the form of English dissertation introduction writing, and puts forward the method of how to write the introduction of the research dissertation by drawing on the good cases. Further, literature [9] analyzed the introduction part of doctoral dissertations by CARS model, and the use of speech steps in these dissertations has less impact on the quality of writing, dissertation writing, and research value. Literature [10] used the CARS model to point out the lack of language steps in undergraduate dissertations, most of which lacked a step in M1, M2, and M3 because of the lack of awareness of the writing process and linguistic resources. These researches, can provide a reference direction to improve the professional writing of thesis and achieve high quality thesis.

In the comparative analysis, literature [11] wrote Indonesian and English dissertation introductions with Indonesian authors based on the CARS model language step, and there is no significant difference between the two in terms of presentation, which shows that the Indonesian authors in the text do not have the same style of writing introductions as the introduction style presented by the CARS model. In contrast, literature [12] analyzed the similarities and differences in the introductions of English language teaching papers and English literature papers using the CARS model, and the results showed that there were differences between the two types of papers in all three areas, M1, M2, and M3. This

study indicates that the three language steps of the CARS model perform differently in terms of whether the dissertation is written in English as a native language or not. Therefore, it is an important direction to study the effect of different language performance on dissertations. In addition, the literature [13] proved the effectiveness of the model in the analysis of abstract genre by analyzing multiple abstract scripts with CARS model, in which all the linguistic steps in the abstracts were clearly identified, and also pointed out that the model can analyze not only the structure of abstracts, but also how to write a rich and professional standard academic paper. And the abstract as the essence of the paper, its language analysis is more research value, but there is a lack of research in this area.

The article discusses the application effect of CARS model based on artificial intelligence in the comparative analysis of medical paper abstracts in English and Chinese. Medical paper abstract information mainly comes from CNKI and Web of Science databases. By incorporating the attention mechanism and global maximum pooling on the Bert model, the learning efficiency of the model on abstract paper extraction is improved. The sigmoid is utilized as the activation function of the output layer of the model for the binary classification problem of abstract information extraction, and the cross-entropy function helps the model to effectively adjust the weights during the training process, so that the output results conform to the real labels as much as possible. The output results of the model are compared and analyzed on the basis of the CARS speech step model to analyze the differences on the Chinese and English abstracts of medical papers.

2. Research on CARS model based on artificial intelligence technology

2.1. Determination of the CARS Discourse Step Analysis Framework

2.1.1. Traditional CARS speech-step modeling. The CARS [14] (Create a Research Space) model was initially used to analyze the introductory section of academic papers, and has since been widely used with the analysis of topics in various writing sections of academic papers. In this paper, this model is used for medical paper abstract genre analysis. The abstract genre analysis framework adopted in this paper is based on the traditional CARS discourse model, and at the same time, it refers to the discourse step analysis frameworks of several academic research literatures. Through the comparative study of theories and the practical operation from the framework to the division of discourse steps, this paper concludes that this framework is more operable. First of all, the traditional CARS step analysis framework is discussed. The traditional CARS framework for discourse step analysis is as follows:

Establish the research area (M1)

Specifically, it includes: identifying the central theme (M1-S1), summarizing the content of the thesis (M1-S2), and reviewing existing research (M1-S3).

Establishing the research position (M2)

Specifically: opposing the argument (M2-S1a), pointing out gaps or lacunae (M2-S1b), posing questions (M2-S1c), and carrying on the research tradition (M2-S1d).

Occupying the research position (M3)

Specifically, this includes: summarizing the purpose of the study (M3-1a), informing about the current research (M3-1b), informing about the main findings (M3-2), and pointing out the structure of the framework of the paper (M3-3).

The traditional framework uses the tokens 1a, 1b and and/or to express the relationship between the steps (S) subdivided under the discourse step (M). It is understood from the literature that all 3 language steps are required, i.e., the absence of any 1 step can be regarded as a poor structural integrity of the introductory language step. The necessity of each step under speech steps 1, 2 and 3 can be categorized into 3 grades from low to high according to their contents, communicative functions, and the combination of markers such as 1A, 1B and and/or.

Grade 1: 4 steps under language step 2, Step 1a: Present a counter argument. Step 1b: Identify gaps or lacunae. Step 1c: Ask questions and Step 1d: Carry forward previous research findings. Step 1a:

Summarize the purpose of the research and Step 1b: Inform about the current research on the topic are set up under language step 3. The above 6 steps are the least necessary 6 steps, and in the original framework, as long as any 2 of the 4 steps under language step 2 occur, language step 2 is considered to be a structurally complete language step. Steps 1A and 1B under step 3 are considered to be structurally complete as long as any one of the two steps occurs.

Level 2: 3 steps under Language Step 1, Step 1: Establish the central theme. Step 2: Summarize the content of the thesis. Step 3: Review existing research. The above 3 steps are the 3 steps of medium necessity, and in Swales' original framework, “Steps 1, 2, and 3”, and the two markers “and/or” are used to indicate that the 3 steps M1-S1, M1-S2, and M1-S3 have a some substitutability.

Level 3: Step 3: "Step 2: Report key findings" and "Step 3: Introduce the structure of the paper". In the original framework of Swales, "steps 2 and 3" were used for labeling, and there were no marks such as "and" or "or" indicating the degree of substitutability. Therefore, M3-S3 and M3-S4 should be included in the required steps.

2.1.2. CARS Discourse Step Analysis Framework Determination. After a discussion of the traditional CARS discourse model, this section will mark the optional steps in the framework with *. On top of the CARS model, it is also worthwhile to refer to the discourse step analysis of medical dissertation abstracts in existing studies.

First, combining the previous studies and the specific corpus of this paper, it is believed that the boundaries of discourse step 1-step 1 and discourse step 1-step 2 are more blurred in the actual academic paper introduction writing, and they should be categorized as optional steps (optional steps are uniformly marked with 1a and 1b, and the serial numbers are followed in order).

Secondly, while the step should be added to the language step 3: summarize the research content of this paper.

Third, the name of step *M2-S1a should be changed to “Demonstrate the value of the study” in the light of the actual corpus. After finalization, the framework of the steps used in this paper is as follows (optional steps are marked with *):

Establishing the research area (M1)

Establishing the central theme (*M1-S1a), summarizing the content of the topic (*M1-S1b), and reviewing existing research (M1-S2).

Establishing the status of the study (M2)

Specifically: arguing for the value of the research (*M2-S1a), pointing out gaps or lacunae (*M2-S1b), posing questions (*M2-S1c), and carrying forward the findings of previous research (*M2-S1d).

Occupying the research position (M3)

Specifically, this includes: summarizing the purpose of the study (*M3-S1a), informing about the current research on the topic (*M3-S1b), outlining the content of the paper's research (M3-S2), informing about the main findings (M3-S3), and describing the structure of the paper (M3-S4).

In summary, there are 4 required steps out of a total of 12 steps set under the 3 language steps. Of the 2 optional steps set under language step 1, the occurrence of any one of them can be considered as a good structural integrity. Of the 4 optional steps set up under language step 2, the occurrence of any one of them can be considered as a good structural integrity of language step 2. In the 2 optional steps under language step 3, again, the occurrence of any of the steps is considered to be better structural integrity.

2.2. Element Extraction Model Construction for Medical Paper Abstracts

2.2.1. Artificial Intelligence Technology Operational Processes. There are many concepts of Artificial Intelligence which are still somewhat controversial among scholars. This paper attempts to understand it to some extent. Artificial Intelligence is a technical science that uses computer systems to study and simulate human intelligence and achieve human-like intelligent behavior. Essentially is

based on big data to achieve the results of training on specific goals, the formation of a certain logical thinking, to achieve the result of getting a high probability of occurrence of behavior, the process of which makes use of self-learning self-improvement of the computer program algorithms, so that the computer shows the characteristics of human behavior, it can be considered that it is actually a large number of human behavioral data collection of the manifestation of the data.

Artificial Intelligence Techniques: advanced forms of statistical and mathematical modeling, such as machine learning, fuzzy logic, and expert systems, usually performed by humans to allow computational tasks. Different AI techniques can be used to implement different AI functions. The flow of AI operations is shown in Figure 1.

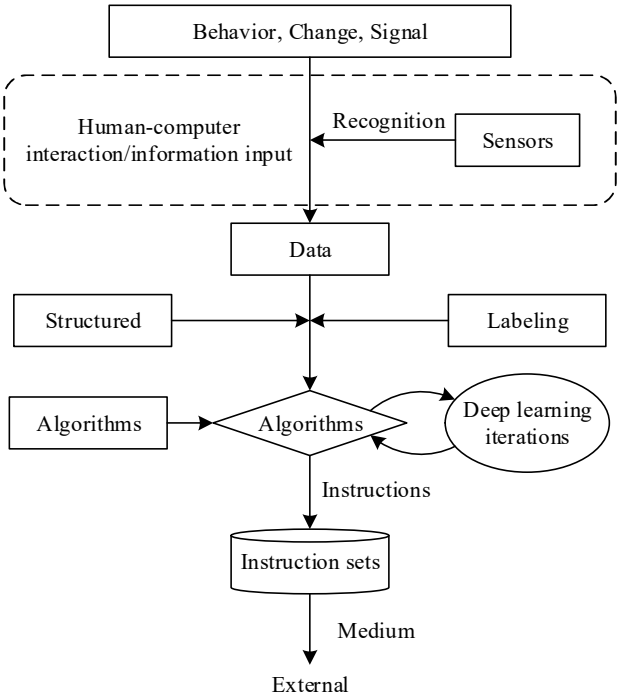

Fig. 1. Artificial intelligence operation process

2.2.2. Model for extracting homogeneous elements of dissertation abstract information. Aiming at the problem of information homogeneity of medical thesis elements, this paper uses element example sentences as restriction items to match thesis sentences, determines whether the thesis sentences contain element semantics, and constructs a two-way input medical thesis homogeneous element extraction model. The abstract of the paper can be obtained through the element extraction of the paper. The abstract extraction of thesis is different from the abstract extraction of general documents. In order to obtain the information homogeneous elements of the abstract, it is necessary to add the upper restriction so that it can focus on the content related to the middle study. However, the sentences containing information about the elements in the paper do not necessarily contain descriptions of medical terms. Therefore, there is a need for a better way of constructing restrictions in medical papers.

Among the textual entailment models, Bert's model [15] had been used in three textual entailment tasks, MNLI (Multi-Genre Natural Language Inference), QNLI (Question Natural Language Inference), and WNLI (Winograd NLI), and they all obtained quite excellent performance. In this paper, based on the Bert model, we constructed a model for extracting homogeneous elements of medical papers. The main architecture of the model is shown in Figure 2.

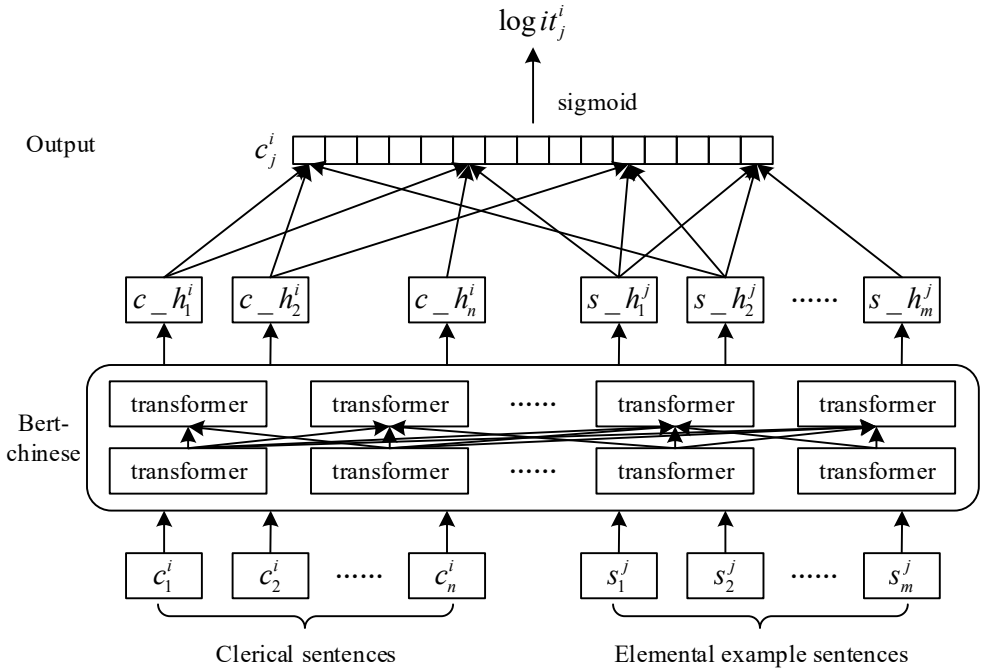

Fig. 2. The main architecture of the model

In this model, for the i st sentence of a medical paper the word level sequence $\left(c_1^i, c_2^i, \ldots, c_n^i\right)$ is matched with the j rd element example sentence $\left(s_1^j, s_2^j, \ldots, s_n^j\right)$.

The matched sentence pairs perform concatenation operations and are then fed into the Bert model to learn word vectors. Inside the Bert model, the input sentence pairs participate in the MLM (Masked Language Model) training task with training objectives such as:

$$P\left(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n\right) \tag{1}$$

That is, using the contextual word sequence $\left(w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n\right)$ of the paper, the probability of position w_i being a particular word is predicted and the generated probability is used as a word vector representation of the character. The final word vector sequence for sentence pairs is obtained as:

$$\left(c\_h_1^i, c\_h_2^i, \ldots, c\_h_n^i, s\_h_1^j, s\_h_2^j, \ldots, s\_h_m^j\right) \tag{2}$$

Also, a single fully connected layer is used as the output layer of the model. The fully-connected layer maps the sequence of word vectors of sentence pairs to the final output probability of the implication relation, e.g:

$$\log it_i = sigmoid\left(W_l \cdot \left[c\_h_i; s\_h_i\right] + b_t\right) \tag{3}$$

2.2.3. A medical paper abstract extraction model incorporating an attention mechanism. Based on the information homogeneous element extraction model of medical thesis, this paper designs a Bleem model, and the architecture of the model is shown in Fig. 3.The main improvements of the Bleem model are:

1) An elemental example sentence is used as a matching term, and the elemental example sentence is used as an auxiliary semantics to match thesis sentences, and a two-way input medical thesis element extraction model is constructed.

2) Using the attention mechanism [16] and global maximum pooling to obtain global information: the attention mechanism of the thesis aligns the element example sentences with the thesis sentences, and

in the process of learning the attention weights, the model can focus on the more relevant parts of the two sentences. Meanwhile, the maximum pooling layer can extract the more obvious hidden layer representation in the connection sequence as the global information for binary classification.

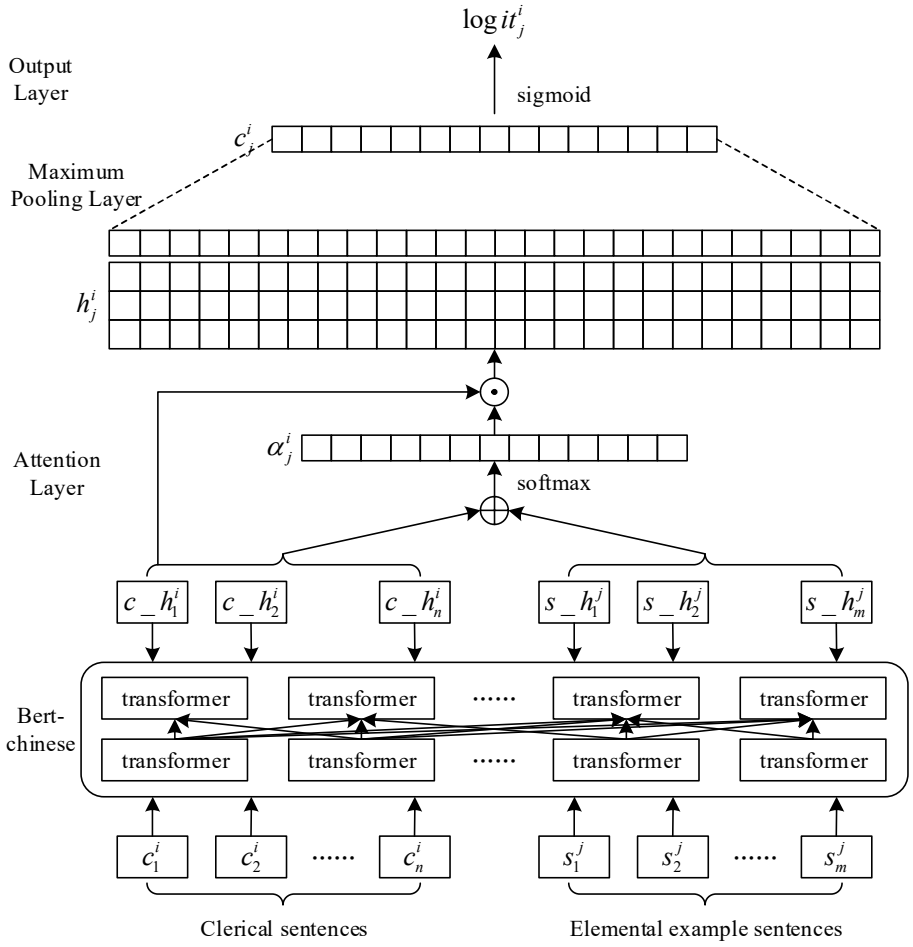

Fig. 3. The model structure of Bleem

Specifically, the i st sentence $(c_1', c_2', \ldots, c_n')$ of the chapter and the j rd example sentence $(s_1^j, s_2^j, \ldots, s_n^j)$ of the set of elements form a sentence pair, where the length of the chapter sentence c_i is n and the length of the elemental example sentence s_i is m. Input to the Bleem model, the Bleem model will process the sentence pairs (c_i, s_i) using a three-layer neural network.

The first layer is the Bert layer, implemented using bert-chinese. Before input to this layer, sentence pairs (c_i, s_i) are concatenated and transformed into three sequences. The first sequence is the word ID sequence, where each word in sentence pair (c_i, s_i) is mapped to a word ID through the lexicon counted from the full corpus. The second sequence is the pre- and post-sentence tagging sequence, where the thesis chapter sentence c_i is transformed into an all-0 sequence of equal length, and the element example sentence s_i is transformed into an all-1 sequence of equal length. The third sequence is a mask sequence with a length equal to the sentence pair (c_i, s_i) as an all-1 sequence. After performing the two learning tasks of word masking and context prediction, Bert outputs a sequence of context-dependent word vectors:

$$\left(c\_h_1^i, c\_h_2^i, \ldots, c\_h_n^i, s\_h_1^j, s\_h_2^j, \ldots, s\_h_m^j\right) \tag{4}$$

The second layer is an attention layer. In this paper, a sequence of word vectors is input into an attention layer, and an attention weight vector is learned based on the sequence of word vectors. After

obtaining the sequence of word vectors of thesis sentence c_h_i and the sequence of word vectors of the elemental example sentence s_h_i, the two sequences of word vectors are connected together and then input into a feed-forward neural network to learn a hidden layer representation z_i^j. Then the hidden layer representation z_i^j is input into $\mathrm{softmax}$, and the values in the z_i^j-vector are mapped to the $(0,1)$-range, and the output is the attentional weights of the thesis sentence for the elemental sentence α_i^j. The refined result is as follows:

$$z_i^j = \tanh\left(W_z \cdot \left[c\_h_i; s\_h_i\right]\right) \tag{5}$$

$$\alpha_i^j = soft\max\left(z_i^j\right) \tag{6}$$

where W_z is the weights of the feed-forward neural network and $\left[c\_h_i; s\_h_i\right]$ denotes the connection between the thesis sentence and the sequence of word vectors of the elemental example sentence.

Using the learned attentional weights α_i of the thesis sentence with respect to the elemental sentence, the sequence of word vectors $c\_h_1^i$ of the thesis sentence is weighted to obtain the rewritten hidden layer representation of the thesis sentence h_i.

$$h_t = \sum_{k=1}^{m} \alpha_{jk}^i \cdot c\_h_k^i \tag{7}$$

where m denotes the length of the thesis sentence, $c\_h_k^i$ is the k rd word representation in the thesis sentence, and α_{jk}^i is the attention weight corresponding to it.

The third layer is the global maximum pooling layer. Due to the weighting, the representations of the more critical words for classification in h_i will be more different from the representations of the non-critical words, and the use of the global pooling layer, which extracts the hidden layer representations with larger values in the hidden layer representations h_i as the representations of the sentences, is more helpful for calculating the implication probabilities of the thesis sentences. Inputting the hidden layer representation h_i of the thesis sentence, the layer will output the pooling vector c_i of the thesis sentence as:

$$c_i = Global\_max\_pooling\left(h_i\right) \tag{8}$$

The fourth layer is the output layer, implemented using a fully-connected layer with a hidden layer unit number of 1. For each thesis sentence, a pooled vector c_i of thesis sentences is input to this layer, which outputs an implied probability vector $\log it_i$ of length 1:

$$\log it_i = sigmoid\left(W_t \cdot c_i + b_l\right) \tag{9}$$

where W_l and b_l are the weights and bias of the output layer, respectively, and sigmoid is used as the activation function of the output layer.

Meanwhile, the loss function of the Bleem model is constructed using the cross-entropy function in the form of:

$$loss = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \log\left(\log it_i\right) + \left(1-y_i\right)\log\left(1-\log it_i\right)\right] \tag{10}$$

Where, y_i is the true value, $\log it_i$ is the predicted value and n is the number of samples.

3. Acquisition and processing of data for medical dissertation abstracts

3.1. Data acquisition

In this paper, Python's Request toolkit is used to crawl the text data of web pages and Lxml toolkit is used to parse the text data in order to realize the acquisition of literature data.

The papers are selected from the Biomedical Literature Service System (BMS), which includes all kinds of Chinese and English biomedical journals as well as doctoral and master's theses since 1978. The search keyword for literature acquisition is medicine, and the time limit of literature is 2020-2024. As more researchers have studied this field in recent years, and the related literature has shown an explosive growth, in order to make the acquired literature more valuable for research, the paper studies three parts of the literature:

The first part is the Chinese journal literature in the field of medicine by the biomedical literature service system.

The second part is the doctoral and master's degree theses in the medical field that are used in the biomedical literature service system.

The third part is the high level medical literature published by Chinese scholars included in the biomedical literature service system.

Based on the above crawling principles, the paper crawled the high-level literature information with medicine as the keyword from 2020 to 2024. Among them, there are 101,965 Chinese journals, 30,502 doctoral and master's degree theses, and 32,619 documents published by Chinese authors in foreign language journals, totaling 165,086 documents.

3.2. Data pre-processing

3.2.1. Data cleansing. Literature was required to delete irrelevant content such as volume headings, book reviews, submission instructions, journal catalogs, announcements, etc. according to the research needs. At the same time, literature that may affect the accuracy of the research results of the paper to a certain extent, such as reviews, research status and prospects, etc., were also deleted.

Basic preprocessing operations were performed on the above data: deletion of literature with missing abstracts or keywords, deletion of duplicates, and standardization of English. The English unification process includes the unification of case, the unification of proper nouns and abbreviations, and stemming extraction. In the case unification process, all upper case English letters are processed into lower case to prevent double counting in the subsequent analysis.

3.2.2. Chinese and English Segmentation Processing. When the literature abstracts are processed for word segmentation, this paper uses jieba lexicon for word segmentation because the third-party library of Python, jieba lexicon, has a higher accuracy rate and can be loaded with custom dictionaries. In this paper, we control the segmentation granularity by emphasizing the high word frequency, and add custom dictionaries to solve the problem by setting the frequency of words in the custom dictionaries. The custom dictionary comes from the high-frequency keywords obtained by frequency statistics after the initial participle processing, the high-frequency keywords are added to the custom dictionary, and the keywords incorrectly participle after the initial participle processing are also added to the custom dictionary, so as to construct a participle library specialized in the field of artificial intelligence.

3.2.3. Discontinued word processing. Discontinued words are conjunctions, auxiliaries, prepositions, etc. that appear frequently in documents but have little practical significance for text analysis. By constructing a list of deactivated words specific to the medical field, the removal of such words of little significance can support the subsequent topic mining in the abstract section of the document.

4. Application of comparative analysis of Chinese and English abstracts of medical papers

4.1. Comparative Analysis of Abstract Sentence Count, Word Count, and Long and Short Sentences

On the basis of the above data corpus, 100 abstracts of Chinese and English medical papers were taken from each of them, and the model of this paper was used to extract and analyze the number of sentences, the number of words, and the number of sentences of long and medium phrases contained in each abstract, and the corresponding percentages were calculated. To provide a research basis for later comparative analysis based on the four common steps of academic papers, i.e. Introduction, Method, Result, Discussion.

Figure 4 shows the results of the comparison of the number of sentences in Chinese and English abstracts. Through the data in the figure, it can be found that in terms of distribution, the number of sentences in English abstracts is concentrated, while the number of sentences in Chinese abstracts is more dispersed. Most of the English abstracts are 6-7 sentences, accounting for 60% of the total. The majority of Chinese abstracts are 2-8 sentences, accounting for 95% of the total. Although the maximum number of sentences in both English and Chinese abstracts is 6, in comparison, English abstracts generally have more sentences than Chinese abstracts. This difference is related to the language steps analyzed below. Chinese abstracts are not standardized in writing, and the abstract language steps are incomplete, while English abstracts have strict requirements and complete language steps.

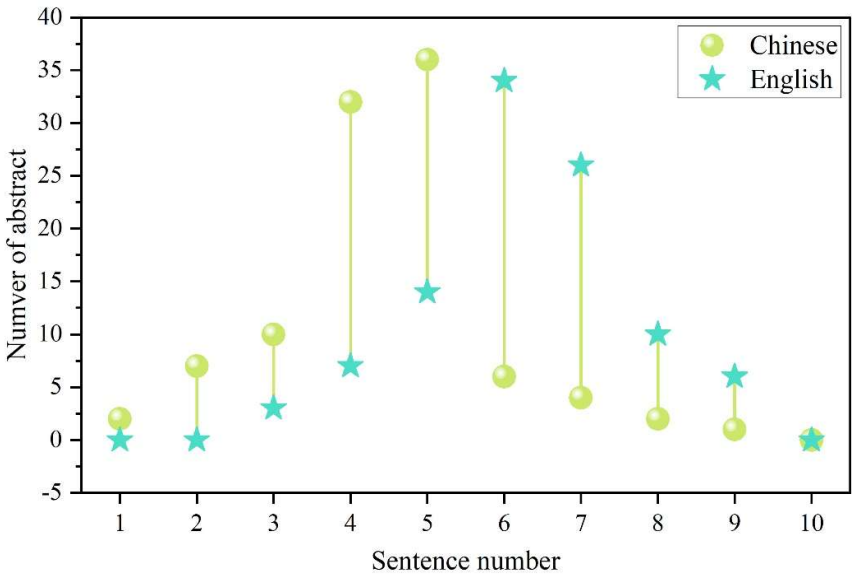

Fig. 4. The comparison of sentence Numbers in Chinese and English

Table 1 shows the results of the word count comparison between Chinese and English abstracts. It can be concluded that Chinese abstracts are generally longer than 350 words, of which 68% of the total number of English abstracts are more than 350 words. On the other hand, most of the English abstracts are less than 400 words, accounting for 96% of the total, but more than 300 words accounted for 61% of the total, indicating that Chinese abstracts are still on the long side compared with English abstracts. This may be due to the fact that Chinese and English are two different languages, with Chinese emphasizing euphemisms and obscurity, while English speakers favor straight-line thinking and cut to the chase.

Table 1. The comparison of words in Chinese and English

Project		Chinese		English	
		Text number	Proportion	Text number	Proportion
Word count	100-200	2	2%	8	8%
	200-250	6	6%	31	31%
	300-350	24	24%	43	43%
	350-400	39	39%	14	14%
	>400	29	29%	4	4%
	In total	100	100%	100	100%

Long and short sentences are relative and there is no certain numerical regulation, in order to facilitate the comparative analysis, it is tentatively decided here that sentences between 1-20 words are short sentences, sentences between 20-40 words are medium sentences, and long sentences are sentences of more than 40 words (not including 40). Figure 5 demonstrates the distribution of short, medium, and long sentences in English and Chinese abstracts as a percentage of the full abstract. The data in the figure show that English abstracts mostly use medium sentences, with an average percentage of 45.24%, while Chinese abstracts mostly use long sentences, with a percentage of more than 80%. Because of the language structure, Chinese focuses on rhetorical flourishes and favors long sentences to express ideas, while English emphasizes simplicity and is accustomed to using fewer words to express more connotations.

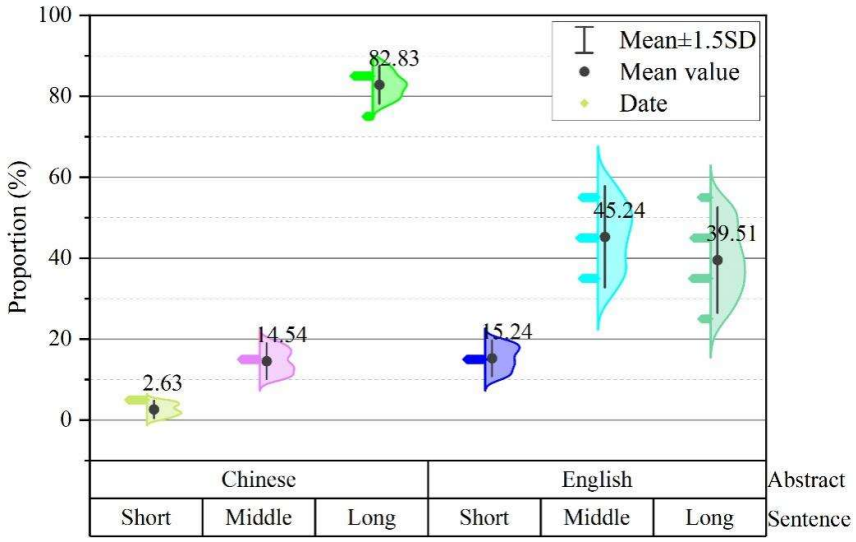

Fig. 5. The proportion of the short, medium, long sentences in the whole abstract

4.2. Grammatical Types of English-Chinese Fuzzy Restrictors in Abstract Language Steps

This section focuses on identifying and analyzing the pragmatic types and fuzzy restrictive phrases of the above English medical paper samples through the model. After the comparison of Chinese and English abstracts, it reveals the overall distribution characteristics of English-Chinese pragmatic types as well as fuzzy restrictive phrases in Chinese-English medical paper abstracts. In order to study the grammatical types of English-Chinese fuzzy restriction phrases in medical abstracts, these types were counted. The data show that the grammatical types of English-Chinese fuzzy restrictive language in abstracts are mainly expressed as, nouns, modal verbs, adjectives, adverbs and verbs. The top 20 words with the highest frequency of fuzzy restriction in 100 abstracts in both Chinese and English are now categorized and counted by grammatical types.

Figure 6 shows the frequency of use of the top 20 words with the highest frequency of fuzzy restrictive phrases in abstracts. From the figure, it can be seen that the frequency of fuzzy restrictive phrases in English abstracts is significantly higher than that in Chinese abstracts. The average

frequency of the first 20 words of fuzzy restrictive phrases in English abstracts is 6328 times, and the highest number of occurrences is “report”, with an average frequency of 498 times. The highest number of occurrences in Chinese is “observation”, with an average frequency of only 98 times. This suggests that the English abstracts tend to use more vague qualifiers to avoid overly absolute expressions, making the study conclusions more rigorous and persuasive.

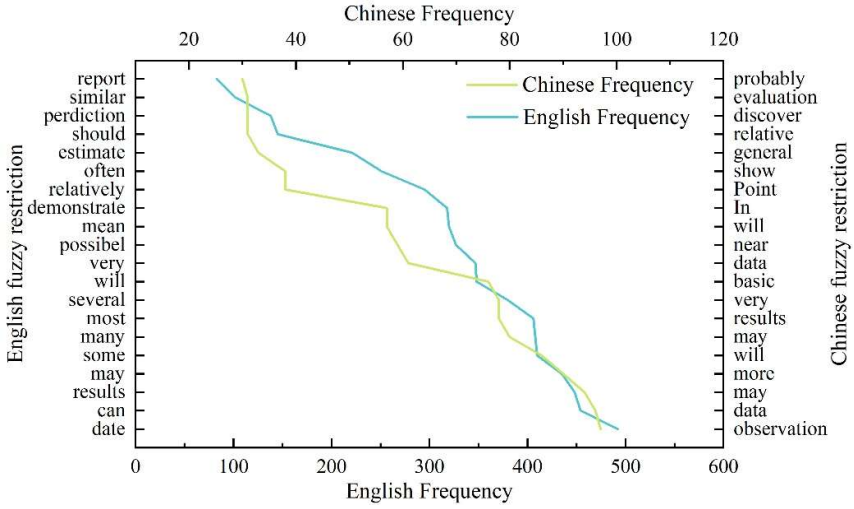

Fig. 6. The first 20 words of the frequency of ambiguity are the first 20 words

The words with the highest frequency in the above figure are categorized by discourse type, resulting in the use of fuzzy restrictors of various grammatical types as shown in Table 2. As can be seen from the table, the use of various grammatical types of fuzzy restrictive phrases in English abstracts is relatively uniform, with a distribution of 14.89% to 25.71%, while the use of adverbs in Chinese abstracts is only 2.42%. It indicates that the authors in Chinese abstracts are more willing to tend to use concise and objective language to describe the results of medical experiments.

Table 2. The use of fuzzy limits of various grammatical types

Fuzzy restriction		English fuzzy restriction		Chinese fuzzy restriction	
		Mean ratio	Error	Mean ratio	Error
Grammatical type	Nouns	17.04%	0.073	18.94%	0.082
	Modal	21.84%	0.085	18.61%	0.058
	Adjective	25.71%	0.063	2.42%	0.059
	Adverb	20.53%	0.078	36.50%	0.073
	Verbs	14.89%	0.076	23.53%	0.063

4.3. Comparison of the number of language steps and the organization of language steps

The abstract is both a summary of the full paper or a condensation of the whole paper and a discourse unit independent of the paper. In terms of discourse cognitive structure, it contains the research question/purpose (Move1), the research methodology (Move2), the main research results (Move3), and the main conclusions (Move4).

In this section, taking Swales' IMRD four-step model as the theoretical framework, based on the communicative purpose and the cognitive structure of language class of each dissertation abstract, and synthesizing its grammatical features and iconic semantic features, the number of steps of abstracts can be divided, and then the model can be used to identify the step structure in each abstract.

Table 3 is obtained to show the results of the analysis of the number of speech steps and structure of Chinese and English abstracts. From the table, it can be seen that there is no case in which the

abstracts of both Chinese and English medical papers contain only one speech step, but there are obvious differences in the distribution of the number of other speech steps. Abstracts of Chinese medical papers generally contain more than two language steps, concentrating on three and four language steps, of which three language steps are mostly used, and the frequency of using two language steps is greater than that of abstracts of English medical papers. The language steps of English medical paper abstracts use a larger proportion of four language steps, accounting for 54%, and are less than the percentage of Chinese medical paper abstracts in both the use of two language steps and three language steps.

In addition, the organization of language steps in the abstracts of Chinese and English medical papers was basically in the order of introduction-methods-results-discussion, but the proportion of such steps in the abstracts of English medical papers was significantly higher than that in the abstracts of Chinese medical papers. The number of abstracts with introduction-methods structure is 9, accounting for 9% of the abstracts, and the structure of three language steps is based on introduction-methods-discussion, accounting for 24% of the abstracts, while the structure of other language steps is only in the order of introduction-methods-results-discussion, accounting for 24% of the abstracts. And there are only 5 articles of other language step organization structure (e.g. Introduction-Results-Discussion, Methods-Results, Methods-Discussion).

The main language step organization structure of English medical paper abstracts was introduction-methods-results-discussion, which accounted for 71%. The number of abstracts with introduction-methods, introduction-methods-results, and introduction-methods-discussion as the language step organization structure was even less than that of English abstracts of Chinese medical papers. Among the three language step organization structure, the abstracts of English medical papers were also mainly based on introduction-methods-discussion, accounting for 19%.

Table 3. Analysis of language steps and structural analysis in Chinese and English

Project		Chinese		English	
		Text number	Proportion	Text number	Proportion
Language step number	One step	0	0%	0	0%
	Two steps	34	34%	17	17%
	Three steps	47	47%	29	29%
	Four steps	19	19%	54	54%
	In total	100	100%	100	100%
Language step structure	Introduction-method	9	9%	2	2%
	Introduction-method-results	8	8%	7	7%
	Introduction-method-discussion	24	24%	19	19%
	Introduction-method-results-discussion	54	54%	71	71%
	other	5	5%	1	1%
	In total	100	100%	100	100%

Comparison of the frequency distribution of language steps in English abstracts of Chinese and English medical papers is shown in Table 4. According to the optimized CARS model of this paper, the number of abstracts of English medical papers is biased at the level of the two language step modes of establishing the scope of the study (M1) and indicating the gap of the research field (M2). Establishing the scope of the study is on the high side of frequency in English medical abstracts, accounting for 45%

of the 100 samples taken, indicating that the English abstracts of English medical papers establish the scope of the study at the beginning.

In terms of M1 (establishing the scope of the study), the abstracts of Chinese medical papers were dominated by summarizing the topic (16%), with a small number of abstract chapters (13% in total) in pointing out the importance of the study and reviewing the results of previous studies. Abstracts of English medical papers were dominated by summarizing the topic (26%), followed by pointing out the importance of the study (17%), and there were only 2 abstracts reviewing previous research results.

In terms of the frequency of the M2 (indicating gaps in the research field) step, 33% of the abstracts of English medical papers included this step, while 20% of the abstracts of Chinese medical papers included this step, which is a significant difference between the two. Both Chinese and English medical abstracts did not include the step of refuting an existing viewpoint in the step of "identifying gaps in the field of research". It is worth noting that neither the English nor the Chinese medical abstracts contain the step "fill in the gaps in the field of research". Overall, the frequency of the CARS step is higher in English medical abstracts, while only half of the Chinese medical abstracts use this step. It can be seen that English medical paper abstracts pay more attention to Swales' IMRD and CARS speech step patterns, which are also more formally relevant.

Table 4. The comparison results of the verbal gait of medical paper

The CARS language step pattern of the abstract		Chinese		English	
		Text number	Proportion	Text number	Proportion
M1	M1-S1a	9	9%	17	17%
	M1-S1b	16	16%	26	26%
	M1-S2	4	4%	2	2%
	In total	29	29%	45	45%
M2	M2-S1a	0	0%	0	0%
	M2-S1b	10	10%	16	16%
	M2-S1c	3	3%	5	5%
	M2-S1d	7	7%	12	12%
	In total	20	20%	33	33%
M3	M3-S1a	0	0%	0	0%
	M3-S1b	0	0%	0	0%
	M3-S2	0	0%	0	0%
	M3-S3	0	0%	0	0%
	M3-S4	0	0%	0	0%
	In total	0	0%	0	0%
In total		49		78	
Proportion		49%		78%	

5. Conclusion

In this paper, based on the previous research, the conventional CARS speech step model was improved and the framework of CARS speech step analysis for medical paper abstracts was identified. In addition, the information of medical paper abstract literature data was collected for the training learning of the medical paper abstract extraction model incorporating the attention mechanism in this paper. The training model and the CARS model were also incorporated into the comparative analysis of medical paper abstracts in English and Chinese, and the strategy of this paper was utilized to explore the differences between culture and language in academic writing. The article obtained the following conclusions with the assistance of the medical dissertation abstract extraction model:

1) The number of sentences in English abstracts is in the range of 6-7 sentences, while the number of sentences in Chinese is scattered in the range of 2-8 sentences, accounting for 60% and 95%, respectively. In addition, the word count of Chinese abstracts is generally more than that of English abstracts, and Chinese abstracts are more inclined to the use of long sentences.
2) The frequency of fuzzy restrictive phrases in Chinese abstracts is lower than that in English abstracts, and the word with the highest frequency only occurs 98 times, and the proportion of adverbs in fuzzy restrictive phrases is relatively small, which is 2.42%.
3) The proportion of English abstracts using four-language steps accounts for 54%, while the use of two-language steps and three-language steps is less than that of Chinese abstracts. In the established research scope of the CARS model, both Chinese and English abstracts mainly summarize the topic, accounting for 16% and 26%, respectively.

By integrating AI technology into the CARS model, the application in the analysis of medical paper abstracts in Chinese and English is more scientific and effective, providing strong support for academic writing. In the future, the AI semantic extraction model can be improved to expand the range of data samples and improve the operability and consistency of the model.

Acknowledgements

The research was founded within the 2013004Srp project entitled Comparative Analysis of Chinese and English Abstracts of Medical Papers Based on CARS Model, which is a key project of Humanities in Scientific Research of Universities in Anhui Province supported by Anhui Education Department.

References

[1] Nurjati, N., & Syahria, N. (2024, August). CONCEPTUAL INSIGHT ON SWALES'CARS MODEL AS AN ACCELERATION OF EFL MASTERS'STUDENTS THESIS WRITING. In International Conference on Language and Language Teaching (pp. 043-054).

[2] Dabamona, M., Satria, D. E., Fatimah, A., & Rahayu, S. (2022). Create a Research Space (CARS) in Introduction of Indonesian Bachelor Thesis: A Corpus-Based Analysis. Journal of English Education and Teaching, 6(1), 126-141.

[3] Dulgheru, A. R. (2022). THE FRAMEWORK OF ACADEMIC ESSAYS SEEN THROUGH A SWALESIAN 'CARS'LENS. Synergies in Communication, 1(1), 196-209.

[4] Bazoubandi, H., & Hajikhani, A. (2023). Communicative-Rhetorical Moves in Research Article Introductions: A Genre Analysis within Swales' Model (2004)(A Case Study of 8 Journals of Quran and Hadith Sciences Genre). Journal of Linguistics & Khorasan Dialects, 15(1), 23-52.

[5] Tullu, M. S. (2019). Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key. Saudi journal of anaesthesia, 13(Suppl 1), S12-S17.

[6] Li, C., Fang, X., Qin, D., & Hua, F. (2022). The structure format of abstracts: a survey of leading dental journals and their editors. Journal of Evidence-Based Dental Practice, 22(3), 101646.

[7] Sánchez, J. A. (2018). Applicability and variation of Swales' Cars model to applied linguistics article abstracts. Elia: Estudios de Lingüística Inglesa Aplicada, 18, 213-240.

[8] Qamariah, H., & Wahyuni, S. (2017). How a Research Article Intruduction Structured? The Analysis of Swales Model (cars) on English Research Article Introductions. Getsempena English Education Journal, 4(2), 136-146.

[9] Hussain, N., Hussain, M. B., & Khan, M. A. (2022). Rethinking PhD Thesis Introductions from Higher Education Perspective: Using the CARS Model. Archives of Educational Studies (ARES), 2(2), 184-204.

[10] Rochma, A. F., Triastuti, A., & Ashadi, A. (2020). Rhetorical styles of Introduction in English language teaching (ELT) research articles. Indonesian Journal of Applied Linguistics, 10(2), 304-314.

[11] Hatmanto, E. D. (2021, December). A comparison between an indonesian and an english journal published in indonesia. In International Conference on Sustainable Innovation Track Humanities Education and Social Sciences (ICSIHESS 2021) (pp. 33-37). Atlantis Press.

[12] Uyman, E. (2017). An investigation of the similarities and differences between english literature and english language teaching master's theses in terms of swales' cars model. PEOPLE: International Journal of Social Sciences, 3(2), 552-562.

[13] Junanto, H., Zahrohtul, G., Triyawan, D., Suminar, R. P., & Wahidah, F. S. (2024). Abstracts Analysis on Student's Final Project Using CARS Theory. Journal of Linguistics, Culture and Communication, 2(2), 190-203.

[14] H U Lemke. (2024). CARS 2025 Computer Assisted Radiology and Surgery - 40th Anniversary and reflections on the role of modelling and AI. International journal of computer assisted radiology and surgery.

[15] Xiao Zunlan & Ning Xiaohong. (2025). Dynamic BERT-SVM Hybrid Model for Enhanced Semantic Similarity Evaluation in English Teaching Texts. Journal of English Language Teaching and Applied Linguistics(1),01-11.

[16] Huapeng Lin,Liyuan Gao,Mingtao Cui,Hengchao Liu,Chunyang Li & Miao Yu. (2025). Short-term distributed photovoltaic power prediction based on temporal self-attention mechanism and advanced signal decomposition techniques with feature fusion. Energy134395-134395.