information

Article

Traditional Chinese Medicine Word Representation Model Augmented with Semantic and Grammatical Information

Yuekun Ma [1,2,3,*], Zhongyan Sun [1], Dezheng Zhang [2] and Yechen Feng [1]

[1] College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; sunzyan@stu.ncst.edu.cn (Z.S.); fengyechen@stu.ncst.edu.cn (Y.F.)
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; zdzchina@ustb.edu.cn
[3] Hebei Provincial Key Laboratory of Industrial Intelligent Perception, Tangshan 063210, China
* Correspondence: mayuekun@ncst.edu.cn

Abstract: Text vectorization is the basic work of natural language processing tasks. High-quality vector representation with rich feature information can guarantee the quality of entity recognition and other downstream tasks in the field of traditional Chinese medicine (TCM). The existing word representation models mainly include the shallow models with relatively independent word vectors and the deep pre-training models with strong contextual correlation. Shallow models have simple structures but insufficient extraction of semantic and syntactic information, and deep pre-training models have strong feature extraction ability, but the models have complex structures and large parameter scales. In order to construct a lightweight word representation model with rich contextual semantic information, this paper enhances the shallow word representation model with weak contextual relevance at three levels: the part-of-speech (POS) of the predicted target words, the word order of the text, and the synonymy, antonymy and analogy semantics. In this study, we conducted several experiments in both intrinsic similarity analysis and extrinsic quantitative comparison. The results show that the proposed model achieves state-of-the-art performance compared to the baseline models. In the entity recognition task, the F1 value improved by 4.66% compared to the traditional continuous bag-of-words model (CBOW). The model is a lightweight word representation model, which reduces the training time by 51% compared to the pre-training language model BERT and reduces 89% in terms of memory usage.

Keywords: word embedding; verb–core structure; traditional Chinese medicine text; part-of-speech; word order; lightweight word representation model

Citation: Ma, Y.; Sun, Z.; Zhang, D; Feng, Y. Traditional Chinese Medicine Word Representation Model Augmented with Semantic and Grammatical Information. *Information* **2022**, *13*, 296. https://doi.org/10.3390/info13060296

Academic Editor: Diego Reforgiato Recupero

Received: 26 April 2022
Accepted: 6 June 2022
Published: 10 June 2022

1. Introduction

Against the background of intelligent medical care, promoting the intelligent development of traditional Chinese medicine (TCM) has become a Chinese government development strategy. The adoption of natural language processing technology to process the TCM text is important for the subsequent intelligent learning of TCM knowledge. Many tasks in the field of natural language processing first require converting words into real-valued vectors, also known as word embedding, which is used as the initial input of downstream tasks. The downstream tasks then choose to extract different feature information for use according to different task requirements. How to use the word representation model to convert as much feature information as possible in the TCM text into the vector space has become the focus of research.

In recent years, many scholars have conducted extensive research on word representation learning methods. Among them, distributed word representation models based on context co-occurrence, such as Word2vec [1,2], GloVe [3], etc., map words into continuous low-dimensional, real-valued vectors. Distributed word representation models improve the effectiveness of one-hot word representation for semantic information representation and

solve the dimensional explosion problem. However, the word vectors generated by such distributed continuous word representation models are relatively independent, with little contextual relevance, and the ability to extract semantic and syntactic feature information embedded in long-sentence text based on a fixed context window is insufficient.

The ELMo model [4] uses bi-directional long short-term memory (Bi-LSTM) to extract the contextual information of the whole sentence by breaking the limitation of the context window and dynamically updating the semantic representation of words according to the context of the sentence. A transformer [5] based on the attention mechanism and location encoding improves language model parallelism and long-distance modeling. Using the multi-layer Transformer architecture and the masked language model (MLM) mechanism, the BERT model [6] can generate word representations that contain rich semantic and syntactic information. Lin et al. [7] suggest that BERT vector representation contains word order information and a hierarchical structure similar to a syntactic tree. Tenney et al. [8] found that BERT embeddings contain part-of-speech (POS), dependency syntax and semantic role information. By executing a characterisation probing operation on MLM, Ettinger et al. [9] investigated the efficacy of each layer vector in BERT to encode word-level, syntactic-level, and semantic-level information, respectively. As the most widely used pre-training model based on the Transformer architecture, BERT has achieved optimal results in many downstream tasks such as text classification [10], entity recognition [11,12], and sentiment analysis [13,14]. Deep pre-training models require large training corpus and long training time due to the large parameter size caused by the complex network structure [15].

To achieve a better representation of the semantic and syntactic features of TCM text, this paper designed a shallow text representation model that is no less than the deep representation model. The major contributions can be summarized as follows:

- We construct a lightweight representation model incorporating a three-layer network structure. The model is divided into two cases: verb-central words and non-verb-central words.
- For verb-central words, this paper extracts the basic syntactic information from the TCM text by formulating nine kinds of sentence meaning characterization rules with verbs as the core. For non-verb-central words, the POS weight coefficient is introduced to generate the POS weight vectors, reflecting the different contributions of words with different POS to sentence meaning understanding in TCM text.
- The model uses convolutional networks to extract word order features from context windows and introduces synonyms, antonyms, and analogous word lists to further improve the representation effect of word vectors on the related semantic information.

2. Related Work

2.1. Verb-Core Structure Theory

Because of the variable grammar and syntax of TCM text, it is difficult to extract syntactic features through grammar and syntax, so we need to find an alternative method to extract syntactic features. Academician Qingli Gao pointed out that the intersection of different natural language syntactic understanding is verbs, so the analysis and understanding of sentence meaning through verbs can be considered [16]. According to θ-theory, Fan proposed a verb–core structure, which contains the three elements of (agentive argument, verb, object argument), and used it to represent the minimal semantics of the sentence. At the same time, he used the combination of several verb–core structures within the sentence to achieve an understanding of the semantics of the whole sentence [17,18]. The verb–core structure can achieve the requirement of extracting the basic syntactic meaning of the sentence; however, it cannot understand the logic of the sentence. Jin [19] introduced the verb valence theory of French linguist Lucien Tesnière on the basis of verb–core structure and proposed the functor theory, which solved the problem that the verb could not be matched to arguments due to a lack of explicit limitations between the verb and its arguments. Verb–core structure theory has achieved good results in the knowledge extraction

of the TCM text, Zhu et al. [20] used key verbs combined with their corresponding arguments to construct semantic relationship rules to achieve semi-automatic matching of the semantic relationships in ancient Chinese medical texts. Babiniotis [21] investigated "the theory of specification based on a verb grammar", which relied on the thesis that human language is logically, cognitively, semantically, and syntactically constructed around a verb. Qian et al. [22] verified the usefulness of verb bias and complementizer cues when dealing with sentences containing temporary ambiguity. Zhou et al. [23] applied the verb-based approach to extract the opinion targets. The authors summarized a variety of right-pointing and left-pointing pairing constructions based on opinion verbs to achieve better opinion object extraction results.

2.2. Word Embedding Methods Based on Part-of-Speech

POS as a fundamental property of natural language plays an important role in the extraction of syntactic information when words are embedded. Liu et al. [24] proposed the part-of-speech word embedding model (PWE) by incorporating POS information as a constraint in the CBOW. In this model, the position-dependent weight matrix is used to model the dependency syntax in the context window. Hu et al. [25] constructed a word embedding acquisition model based on the principle of separate learning of nouns and verbs. This model makes the nearest neighbors of word vectors more reasonable based on the assertion that humans use different methods for nouns and verbs learning. Pan et al. [26] used the POS association matrix and the fixed syntactic relationship between words to enhance the quality of the word vectors generated by the Word2vec. Wang et al. [27] introduced POS tags in the field of visual question answering to strengthen the role of the important words and realized the capture of semantic information between the questions and answers based on the vector representation. Deng et al. [28] introduced POS vectors to solve the problem that part-of-speech similarity is difficult to calculate and proposed the CBOW+P+G model. The model makes better use of POS correlation of context to predict central words, and uses dependency syntactic weights to reduce the information loss caused by sliding windows. Ren et al. [29] constructed the attention enhanced Chinese word embeddings (AWE) model by introducing the self-attention mechanism and the position function to solve the problems of the same weight of contextual word mapping and the lack of word order information in the CBOW model.

3. Model Introduction

In this paper, we propose a Verb as Core and Part-of-speech and Convolution Word Embedding (VCPC-WE) model based on the valence verb–core structure, which combines POS and word order information, to achieve TCM word representation that retains more semantic and syntactic information.

3.1. Semantic Model Construction of TCM Text

Language feature modeling based on the analysis of TCM text features.

TCM domain text contains a large number of verbs and fixed sentence patterns [30], such as 为 (pertain), 主 (govern), 恶 (detest), 生 (generate), etc. Many of the sentences are composed of fixed sentences with the above verbs as the core, such as "南方生热，热生火，火生苦，苦生心，心主舌。其在天为热，在地为火，在体为脉，在脏为心，在色为赤，在音为徵，在声为笑，…". In this paper, we analyze and characterize the meaning of the sentence by using the fixed sentence structure with verbs such as 生 (generate) and 主 (govern) as the core. This paper constructed a lexicon containing 105 typical verbs, in which verbs are mainly divided into two categories:

- Words themselves commonly used as verbs.
- Words commonly used as nouns, and there are more cases of nouns being used as verbs in TCM text.

This paper divided the head words of predicted verbs into monovalent verbs and bivalent verbs according to the number of terms constrained by the verbs, and summarized nine kinds of sentence meaning representation rules with verb–core, as shown in Table 1.

Table 1. Verb–core sentence meaning representation rule.

Value of Verbs	Verb Examples	Rules
Bivalent verb	主 (govern) 伤 (impair) 发于 (occur in) 入通于 (related to)	Noun–Verb–Noun Verb–Noun–Noun Noun–Noun–Verb Pronoun–Verb–Noun Noun–Verb–Pronoun Verb–Noun–Pronoun Verb–Pronoun–Noun
Monovalent verbs	至 (reach) 溢泻 (emit)	Noun–Verb Verb–Noun

TCM text is concise and the sentences are mostly compounded by short sentences. Words with different POS have different contributions to the semantic understanding of sentences. For example, 之 (this) and 而 (and) in "肾者主水，受五脏六腑之精而藏之，故五脏盛，乃能泻" are weak semantic words that should be given less attention than strong semantic words such as 主 (govern) and 五脏 (five internal organs). In this paper, we divide the words in sentences into two categories: strong semantic words such as nouns and verbs and weak semantic words such as pronouns and conjunctions. The detailed classification and descriptions are shown in Table 2.

Table 2. Classification of strong and weak semantic roles of words.

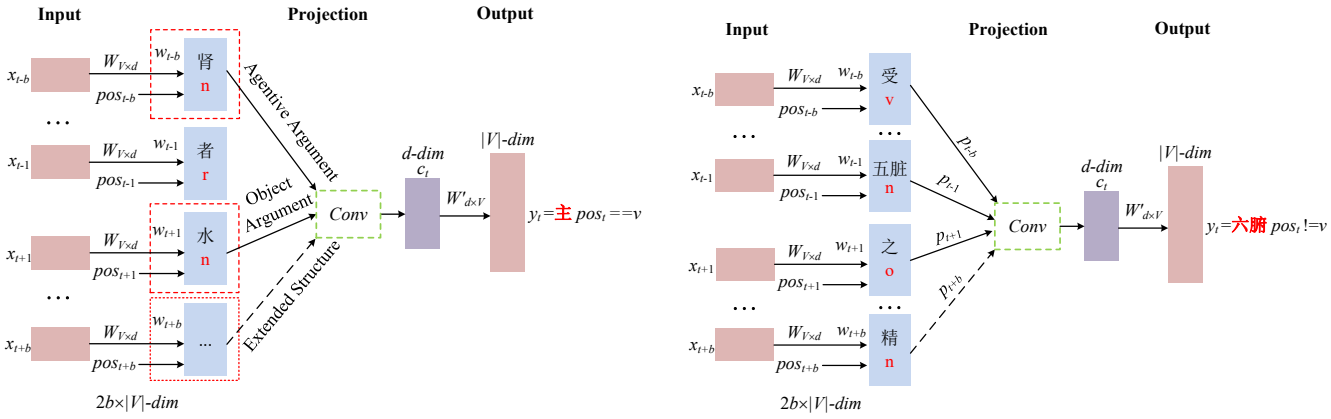
Category	Lexical Category	Description
Strong semantic words	Nouns (n)	Includes person nouns, place nouns, location nouns, and time nouns.
	Verbs (v)	Verbalization of nouns exists after the inclusion of corpus correction.
	Adjectives (a)	
	Quantifier (q)	Includes number and measure words.
Weak semantic words	Pronouns (r)	Includes personal pronouns, interrogative pronouns, etc.
	Adverbs (d)	Includes general adverbs as well as negative adverbs.
	Conjunctive (c)	
	Other lexical (o)	Includes auxiliary words, onomatopoeia, and others.

Different combinations of words in TCM text represent different semantics. For example, "推而外之，内而不外，有心腹积也" and "推而内之，外而不内，身有热也" contain the same words, but the order in which results in different semantic meanings. This paper uses an N-gram convolutional network to extract word order features from the context of the central word.

3.2. VCPC-WE Model Construction

The VCPC-WE model includes a three-layer neural network of input, projection, and output. The target word predicted by the model is the central word, and the words in the windows on both sides of the central word are the context. The input layer is the

context word vector of the central word and the corresponding POS tags. The projection layer adopts two different intermediate vector calculation methods according to whether the central word is a verb, corresponding to the calculation of c_t in Figure 1a,b, respectively. In addition, the projection layer extracts word order features through convolution operations, and the output layer uses intermediate vectors and parameter matrices to predict the center word. The model is divided into two cases: verb-central words and non-verb-central words, and the overall framework is shown in Figure 1.



(**a**) Model structure based on verb-central word

(**b**) Model structure based on non-verb-central word

Figure 1. Overall framework of the VCPC-WE model. Based on whether the central word y_t is a verb, the associated word vectors of central word vector w_t are obtained according to (**a**,**b**). The contextual intermediate vector c_t is computed by the convolution operation *Conv* for the obtained associated vectors. The VCPC-WE model uses the vector c_t with all words in the word list to calculate the conditional probabilities and to maximize the conditional probability $P(y_t|c_t)$ for the central word.

The embedding matrix of the input layer of the VCPC-WE model is denoted as $W \in \mathbb{R}^{|V|\times d}$, where d is the word vector dimension, and $|V|$ is the size of the training vocabulary $V = \left\{v_1, v_2, \ldots, v_{|V|}\right\}$. The mapping matrix of the output layer is denoted as $W' \in \mathbb{R}^{d\times|V|}$. x and y denote the one-hot encoding of the word v, and w denote the word embedding of the word v; thus, $w = Wx$. To predict the central word vector w_t, VCPC-WE uses a sequence of contextual word vectors $H = \{w_{t-b}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+b}\}$ and corresponding POS tag sequences $POS = \{pos_{t-b}, \ldots, pos_{t-1}, pos_{t+1}, \ldots, pos_{t+b}\}$ to generate an intermediate vector c_t, where hyper-parameter b represents the context window size.

3.2.1. Model Solution Based on Verb-Central Word

When the central word is a verb, the VCPC-WE model uses the combination rules in Table 1 to match the corresponding agentive argument and object argument according to the cost of the word and the POS tag of the word. If the determiner of the agentive and object argument (such as the adjective corresponding to the word) exists in the context, it is used as an extended structure to characterize the basic semantics of the sentence. For example, when the central word is 主 (govern), the corresponding agentive argument in the context window is 肾 (kidney), the object argument is 水 (water), there is no extended structure in the context. The central word y_t is predicted by VCPC-WE by calculating the conditional probability $P(y_t|Context(y_t))$ of the central word and the context word $Context(y_t)$:

$$P(y_t \mid Context(y_t)) = P\left(y_t \mid w_{\{t-b,\ldots,t-1,t+1,\ldots t+b\}}\right) = \frac{\exp\left(y_t^T W' c_t\right)}{\sum_{x\in V}\exp(x^T W' c_t)} \quad (1)$$

The subscript set of the agentive argument is defined as AA, the object argument is OA, the extended structure is ES, and the middle vector c_t is calculated as

$$\begin{aligned} c_t &= \sum_{j \in \{t-b,\ldots,t-1,t+1,\ldots,t+b\}} v\left(Context(y_t)_j\right) \\ &= \frac{1}{n} \sum_{j \in \{AA,OA,ES\}} w_j \end{aligned} \tag{2}$$

where n is the number of lexical items that can be constrained by this verb-central word.

The objective function of the model is to maximize the log-likelihood function L:

$$L = \sum_{y_t \in V} \log P(y_t \mid Context(y_t)) = \sum_{y_t \in V} \log P(y_t \mid c_t) \tag{3}$$

The negative sampling algorithm is used to train the objective function. For the context $Context(y_t)$, y_t is the positive sample, and the remaining words are randomly sampled to form the negative sample set $NEG(y_t)$. Then, the positive and negative samples are labeled as follows:

$$l_y = \begin{cases} 1, y = y_t \\ 0, y \neq y_t \end{cases} \tag{4}$$

In the case of a negative sampling strategy, the calculation of Equation (1) is substituted into Equation (3) to form the objective function of VCPC-WE:

$$\begin{aligned} L &= \sum_{y_t \in V} \log \prod_{y \in \{y_t\} \cup NEG(y_t)} \left\{ \left[\sigma\left(c_t^T \theta^y\right)\right]^{l_y} \times \left[1 - \sigma\left(c_t^T \theta^y\right)\right]^{1-l_y} \right\} \\ &= \sum_{y_t \in V} \sum_{y \in \{y_t\} \cup NEG(y_t)} \left\{ l_y \log\left[\sigma\left(c_t^T \theta^y\right)\right] + (1 - l_y) \log\left[1 - \sigma\left(c_t^T \theta^y\right)\right] \right\} \end{aligned} \tag{5}$$

where y denotes the words in the positive and negative samples, θ^y is the parameter vector calculated from y and the output parameter matrix W', and $\sigma\left(c_t^T \theta^{y_t}\right)$ is the probability of predicting the positive class y_t, calculated as follows:

$$\sigma(c\theta) = \frac{1}{1 + e^{-c^\tau \theta}} \tag{6}$$

To facilitate the derivation of the formula for the gradient calculation, the formula for gradient calculation involving the intermediate vector c_t is abbreviated to $\zeta(c, y)$ in the braces in Equation (5):

$$\begin{aligned} \frac{\partial \zeta(c,y)}{\partial c_t} &= \frac{\partial}{\partial \theta^y} \left\{ l_y \log\left[\sigma\left(c_t^T \theta^y\right)\right] + \left[1 - l_y\right] \log\left[1 - \sigma\left(c_t^T \theta^y\right)\right] \right\} \\ &= l_y \left[1 - \sigma\left(c_t^T \theta^y\right)\right] \theta^y - \left[1 - l_y\right] \sigma\left(c_t^T \theta^y\right) \theta^y \\ &= \left\{ l_y \left[1 - \sigma\left(c_t^T \theta^y\right)\right] - \left[1 - l_y\right] \sigma\left(c_t^T \theta^y\right) \right\} \theta^y \\ &= \left[l_y - \sigma\left(c_t^T \theta^y\right) \right] \theta^y \end{aligned} \tag{7}$$

To update the vector of the agentive and object argument of the central word constraint, the gradient descent method is used:

$$w_i = w_i + \eta \frac{\partial \varsigma(c,y)}{\partial c_t} \quad , i \in \{AA, OA, ES\} \tag{8}$$

where η is the learning rate.

3.2.2. Model Solution Based on a Non-Verb-Central Word

In addition, when the central verb is a non-verb, the input layer consists of a sequence of contextual word vectors H and the corresponding sequence of part-of-speech labels POS. According to Table 2, this paper sets different part-of-speech weight coefficients k_{pos_t} for different part-of-speech tags pos_t based on strong and weak boundaries and uses the initial part-of-speech vector $v(pos_t)$ and the part-of-speech weight coefficient to obtain the part-of-speech weight vector $p_t = k_{pos_t} \cdot v(pos_t)$.

The context vector c_t is calculated as

$$\begin{aligned} c_t &= \sum_{j \in \{t-b,\ldots,t-1,t+1,\ldots,t+b\}} v\left(Context(c_t)_j\right) \\ &= \frac{1}{2b} \sum_{j \in \{t-b,\ldots,t-1,t+1,\ldots,t+b\}} w_j \cdot p_j \end{aligned} \tag{9}$$

The objective function in the case of non-verb-central words is also Equation (3), and the gradient calculation formula for solving the objective function still uses Equation (7). Different from the case of the verb-central word, the model uses different weight coefficients for the backward update of the POS weight vector:

$$\begin{aligned} w_i &= w_i + \eta \frac{\partial \zeta(c,y)}{\partial c_t} \\ p_i &= p_i + k_{pos_i} \eta \frac{\partial \zeta(c,y)}{\partial c_t} \end{aligned} \quad , i \in \{t-b,\ldots,t-1,t+1,\ldots,t+b\} \tag{10}$$

3.2.3. Word Order Feature Extraction

One-dimensional convolution, also known as sequence convolution, has the core function of extracting local relevance features of data, and because word order features denote local relevance between words, one-dimensional convolution can be used to extract sentence word order features. Based on one-dimensional convolution for words in the context window, we propose using N-Gram convolution to preserve word order features, and the model structure is shown in Figure 2.

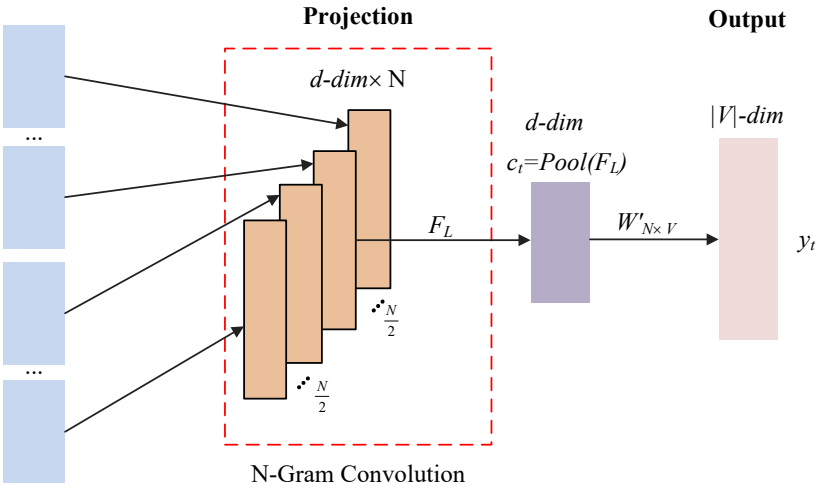


Figure 2. The VCPC-WE convolutional structure figure is an expansion of the *Conv* module in Figure 1. The contextual word vectors which are strongly associated with the central word y_t are extracted to form the context matrix, which contains N d-dimensional word vectors.

N-gram convolution is a convolution operation on a sequence of words with window size N on a sequence of sentences to extract word order features, and the detailed process is shown in Figure 3.

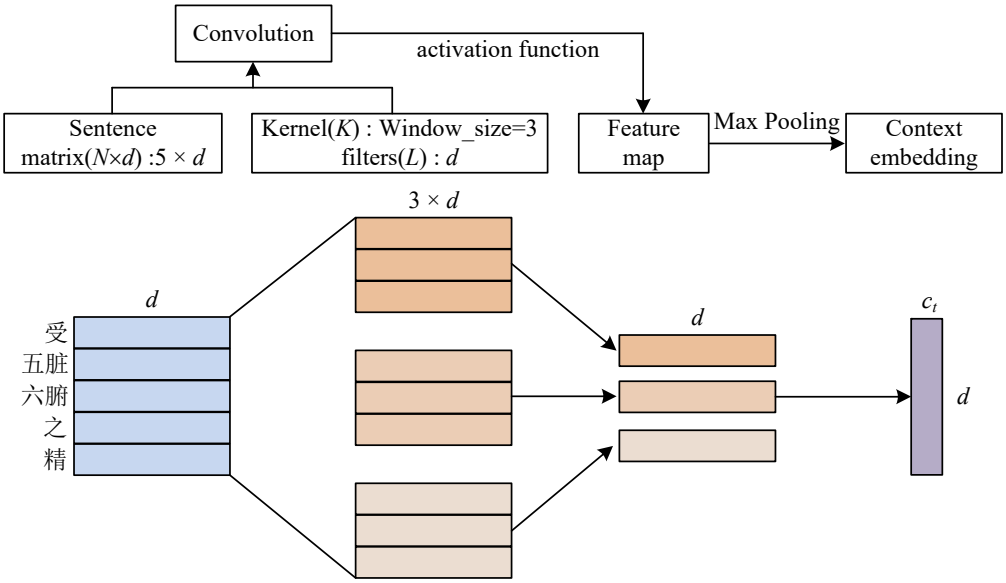


Figure 3. N-gram convolution flow chart, the word vector matrix of size $N \times d$ is convolved by random initialization of L convolution kernels K, then multiple feature vectors are obtained by activation function, and finally the context vector c_t is obtained by max pooling operation.

The input matrix of N-gram convolution is $m \times d$, where m is the number of words in the window to be processed and d is the dimension of the word vector. The appropriate step size parameter K, i.e., the width of the convolution kernel, and the number of convolution filters L are used to perform N-gram convolution. We use the filters to move down the text sequence to the bottom of the sequence to obtain $N - K + 1$ vectors after convolution, and then use max pooling to obtain the local feature information of the context. For the sequence of contextual word vectors H, each contextual word vector w is processed using the convolution kernel K. The convolver $F_L = KH_L$ is obtained, where $K \in \mathbb{R}^{N \times d}$ is the convolution kernel, N is the width of the convolution kernel, and H_L is the sequence of contextual word vectors traversed by the Lth filter.

By using a max pooling operation, the set of all convolution filters F_L is applied to the context window of the central word, yielding the filter set $F = \{F_1, F_2, \ldots, F_L\}$ and then the new intermediate word vector $c_t = Maxpool(F)$.

3.2.4. Synonymy, Antonymy, and Analogy Information

According to the analysis of synonymy and antonymy phrases in the Neijing Language Study [31] and the study of analogy relationship by Wang [32], we constructed a lexicon of synonyms and antonyms (see Table 3), as well as an analogies dictionary (see Table 4).

Table 3. Some synonyms and antonyms.

Synonyms		Antonyms	
神–明 (God–Ming)	征–兆 (Sign–Omen)	本–末 (Begin–End)	长–短 (Long–Short)
肇–基 (Start–Base)	变–化 (Change–Conversion)	表–里 (Surface–Inside)	白–黑 (White–Black)
魂–魄 (Soul–Spirit)	津–液 (Saliva–Fluid)	春–秋 (Spring–Autumn)	迟–数 (Late–Frequent)
懈–惰 (Slack–Lazy)	空–穴 (Empty–Cave)	日–暮 (Sunrise–Sunset)	丑–善 (Ugly–Good)
移–易 (Shift–Change)	分–别 (Divide–Leave)	德–过 (Morality–Fault)	粗–细 (Coarse–Thin)

Table 4. List of some analogous words.

Analogous Phrases			
五脏–五行 (Five Internal Organs–Five Elements)	肺–金 (Lung–Gold) 肝–木 (Liver–Wood) 肾–水 (Kidney–Water) 心–火 (Heart–Fire) 脾–土 (Spleen–Earth)	五脏–五味 (Five Internal Organs–Five Tastes)	肺–辛 (Lung–Pungent) 肝–酸 (Liver–Acid) 肾–咸 (Kidney–Salty) 心–苦 (Heart–Bitter) 脾–甘 (Spleen–Sweet)
四经–四时 (Four Canons–Four Seasons)	肝–春(Liver–Spring) 心–夏(Heart–Summer) 肺–秋(Lung–Autumn) 肾–冬(Kidney–Winter)	阳–天 (Yang–Sky) 阳–日 (Yang–Day) 阳–火 (Yang–Fire) 天气–雨 (Tianqi–Rain)	阴–地 (Yin–Earth) 阴–月 (Yin–Moon) 阴–水 (Yin–Water) 地气–云 (Diqi–Clouds)

For the synonyms in the text, such as "盛–满" in "以耗散其真，不知持满" and "女子七岁，肾气盛，齿更发长", the distance between sentences in the corpus is relatively long. These synonym vectors generated by the shallow word representation model are far away in the vector space and cannot reflect the similarity of synonyms well. In this paper, using the synonym table as a guide, we adopt a two-word averaging strategy for the synonym word vectors of the output matrix W' in the VCPC-WE model backward update generation word vector stage to correct the model, so that the model can calculate a closer spatial distance for the synonym word vectors and preserve the semantic relationship between synonyms:

$$w_i = w_j = \frac{w_i + w_j}{2} \tag{11}$$

In the case of antonyms, the strategy of reversing two words is used to generate the antonymic word vector separately:

$$w_i = -w_j \tag{12}$$

The analogical relationships in TCM text are characterized by a hierarchical structure as well as multiple analogies [33]. For example, the top-level concept words 五脏 (five internal organs) are analogous to the 五行 (five elements), and the sub-level concepts within the five internal organs, such as 心 (heart), 肝 (liver), and 脾 (spleen), are analogous to 火 (fire), 土 (earth), and 木 (wood), respectively. An example of multiple analogy is that the five organs are analogous to both the five colors and the five tastes. Combined with the analogy word Table 4, the VCPC-WE model reverses the update of the output matrix W' by adopting the strategy of averaging the hierarchical and multiple analogous words. The correction formula is:

$$w_{\text{五脏}} = \frac{w_{\text{五脏}} + w_{\text{心}} + w_{\text{肝}} + \ldots + w_{\text{黄}}}{n} \tag{13}$$

where n is the number of all analogous lexical items corresponding to the five internal organs.

4. Experiment and Evaluation

4.1. Data Pre-Processing

In this paper, we choose *Huangdi Neijing* as the object of experimental research. First, we remove the garbled code, unify the punctuation symbols to Chinese symbols, convert traditional Chinese to simplified Chinese, and convert the common characters, such as "五藏" to "五脏". Following data cleaning, the text was divided into sentences based on sentence endings, and a custom word list was created. The word list contains domain phrases compiled by the subject group as well as Chinese medicine symptoms and prescriptions from the Sogou and Baidu cellular thesauri. The sentence was then segmented, and POS tagging was performed using the word segmentation tool LTP. The results of automatic POS tagging include 17 POS with some tagging errors. This paper focuses on parts of speech that are crucial in semantic representation. According to the characteristics of TCM text and literature [27], 17 POS were classified into eight types, including nouns, verbs,

and adjectives, etc. Finally, 10,164 sentences were obtained, as well as 7677 words after de–duplication.

4.2. Experimental Environment Configuration

The model experiments in this paper, as well as the model comparison experiments, were carried out in the same experimental environment, as described below:

- Operating system: Ubuntu 18.04.
- GPU: RTX2080ti, 32G RAM.
- Running framework: Pytorch 1.6.0.

4.3. Experimental Parameter Settings

There are many sentences composed of short sentences in the *Huangdi Neijing*, the context window size of the model is set to 6, and the word vector dimension is set to 100. To optimize the training process, the model employs a negative sampling strategy with five negative samples and a subsampling threshold of 10^{-3}. The BatchSize parameter was set to 120, and the initial learning rate was set to 0.01.

4.4. Evaluation

First, the VCPC-WE model is compared and evaluated with the deep pre-training model in terms of model scale. Second, there are two methods for evaluating the quality of word vectors: internal evaluation and external evaluation. Internal evaluation consists primarily of word relevance and word analogy evaluation. External evaluation is the process of using trained word vectors as input to downstream tasks and evaluating the quality of the word vectors by analyzing the results of these tasks. As an internal evaluation method, we randomly selected a group of semantically related words from the existing word correlation dataset to calculate the corresponding similarity. The word vectors were used as input features for the Huangdi Neijing entity recognition task, and the merits of the word vectors were quantitatively analyzed by comparing entity recognition results.

4.4.1. Model Scale Comparison

VCPC–WE was compared with two types of pre–training models in terms of training time, model parameter size, and resource consumption. The results are shown in Table 5.

Table 5. Model scale statistics.

Model	Training Time (Iter/s)	Parameters (MB)	Memory Usage (MB)
ELMo [4]	17.68	93.6	10,141
BERT [6]	15.3	110	13,631
VCPC-WE	7.52	67.9	1430

Each comparison model was trained with the same batch size and number of epochs, and the training time was compared by comparing the running time of one iteration of each model. The average number of iterations across multiple epochs was used to calculate iteration time. According to Table 5, the VCPC–WE proposed in this paper reduces training time by approximately 51% when compared to BERT and approximately 57% when compared to ELMo. BERT is approximately 14% faster than the ELMo model, demonstrating that the Transformer architecture is more efficient than the LSTM architecture.

The ELMo model saves 15% and the proposed model in this paper saves about 38% when compared to the BERT model with a 12-layer Transformer structure and total parameters of 110 (MB). Because the number of network layers is reduced, the parameter amount and the resource occupation are both decreased. When compared to BERT, the VCPC–WE model reduces the memory footprint by approximately 89%. As a result, the VCPC–WE model significantly reduces time and resource costs, and these benefits make the model more suitable for application to practical tasks.

4.4.2. Qualitative Analysis

By using random numbers, "Yang" was chosen as an experimental reference word from the existing vocabulary related to word meaning in the field of TCM in this paper.

The cosine similarity between the words was calculated using a set of words related to the semantic meaning of "Yang", such as "Sky", "Day", and "Fire", and the antonym "Yin". To determine the effectiveness of the word vectors generated by each model in characterizing semantic information, the cosine similarity between the words was calculated, and the experimental results are shown in Table 6.

Table 6. Case study of word cosine similarity on word pairs about "Yang".

Models	阳 (Yang) and 天 (Sky)	阳 (Yang) and 日 (Day)	阳 (Yang) and 火 (Fire)	阳(Yang) and 阴 (Yin)
CBOW [1]	0.04227	0.02883	0.03077	0.17383
Skip–gram [2]	0.04093	0.03595	0.03356	0.19334
Glove [3]	0.04181	0.03082	0.03682	0.19267
PWE [24]	0.06341	0.04763	0.04965	0.17768
CBOW+P+G [28]	0.06840	0.06167	0.06475	0.17028
AWE [29]	0.07565	0.07203	0.07114	0.15031
ELMo [4]	0.09546	0.07043	0.07908	0.14108
BERT [6]	0.11756	0.15122	0.10916	0.07883
VCPC–WE	**0.19083**	**0.22339**	**0.13990**	**0.04327**

The values in bold represent the best results for all models.

The first three columns are the near-sense word pairs of "Yang", and larger values indicate higher similarity; the last column is the antonym pair of "Yang", and smaller values indicate lower similarity. In terms of the overall ability to obtain semantic information, the experimental results show that the proposed model outperforms the comparison model.

4.4.3. Entity Identification Experiments

Entity recognition is the task of labeling noun concepts in sentences. In this paper, the entities in Huangdi Neijing are divided into four parts: physiological, pathological, cognitive, and syndrome differentiation and treatment. The corpus was labeled with categories using the BIOES labeling system and divided into training, validation, and test sets in a 6:2:2 ratio. The model proposed in this paper was compared to other word embedding models, and the experimental results of each model were obtained by taking the mean value of five experiments, as shown in Table 7.

Table 7. Identification results.

Models	Accuracy Rate %	Recall Rate %	F1 Value %
Skip–gram [2]	83.17	76.09	79.47
Glove [3]	82.07	76.58	79.22
CBOW [1]	83.61	78.30	80.86
PWE [24]	83.24	80.53	81.85
CBOW+P+G [28]	83.92	80.34	82.09
AWE [29]	84.56	81.94	83.23
ELMo [4]	85.24	80.80	82.95
BERT [6]	84.97	82.42	83.67
VCPC–WE	**86.01**	**85.04**	**85.52**

The values in bold represent the best results for all models.

Table 7 shows that the VCPC–WE model proposed in this paper achieves the best recognition effect, with a 4.66% improvement in F1 value over the CBOW baseline model. Since the lack of unified and fixed grammar and syntax in the ancient Chinese texts, the CBOW+P+G model proposed by the literature [28] did not perform well on the text of TCM. Compared with the AWE model, the VCPC-WE model improved the recognition

results by 2.29%. The AWE model enables central words to capture contextual information at longer distances, but there are more short sentences in the TCM text. In addition, attention is given to non-important components, which diminishes the contribution of subject components (nouns, verbs, etc.) to semantic representation. Compared with the deep pre-training models ELMo and BERT, the entity recognition F1 value of the VCPC-WE model proposed in this paper is improved by 2.57% and 1.85%, respectively. The experimental results verify that the VCPC-WE model has the semantic feature extraction ability in TCM text that is not weaker than the deep pre-training model.

A comparison experiment was carried out using the control variables method between the enhancement model VCP–WE, which only incorporates kernel structure and lexical information, and the enhancement model Conv–WE, which only considers word order information (see Figure 4). Compared with the baseline model CBOW, the recognition results of each enhancement method were improved by 2.38% and 1.01%, respectively. In addition, compared with Conv-WE, the recognition effect of VCP-WE is improved by 1.37%, which verifies the effectiveness of sentence meaning representation rules and part-of-speech features in improving the quality of word vectors.

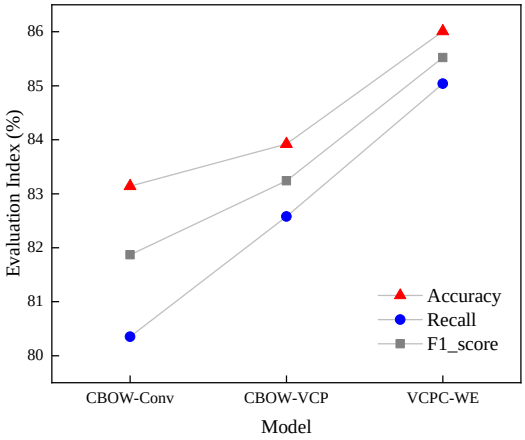


Figure 4. Results of ablation experiments.

4.4.4. Effect of Different Dimensional Word Vectors on the Results

We compared the impact of word vectors of different dimensions on the entity recognition task, and the experimental results are shown in Figure 5, with the resultant data derived by averaging five experiments.

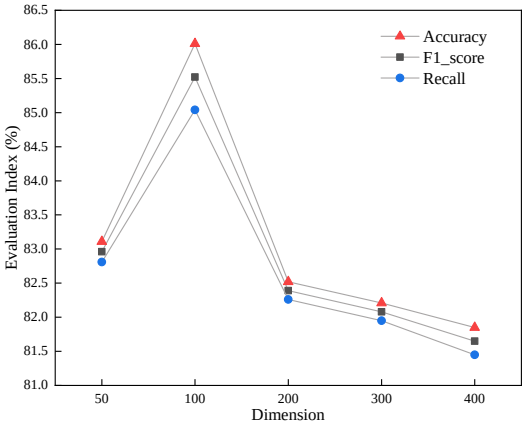


Figure 5. Entity recognition results for word vectors with different dimensions.

Because of their low dimensionality, 50-dimensional word vectors cannot fully characterize semantic information. Due to the sparse semantic information caused by a higher

dimensionality, recognition results for 200-dimensional and higher word vectors are decreasing. Therefore, the VCPC-WE uses 100-dimensional word vectors.

4.4.5. Effect of Different Window Sizes in the Context on the Results

This section compares and analyzes the impact of word vectors generated under various window size parameters on entity recognition tasks.

As shown in Figure 6, when the window size changes from 6 to 8, the F1 value decreases significantly and increases slowly from 8, but its maximum value is still obtained when the window size is 6. Because there are more short sentences in Huangdi Neijing, the window size was set to 6 in this study.

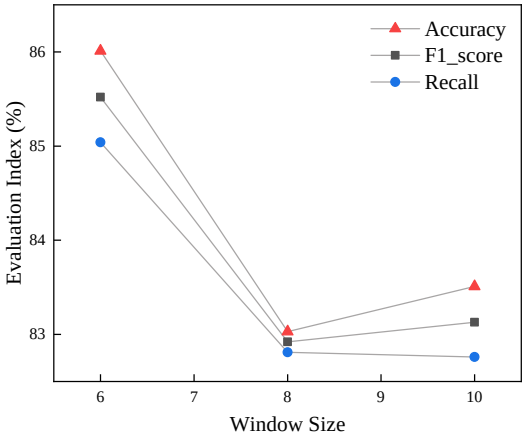


Figure 6. Entity recognition results for word vectors with different windows.

5. Conclusions

This paper proposes a VCPC-WE model based on the linguistic characteristics of the TCM text and the theory of verb–core structure, combined with part-of-speech information. The model enhances the ability to extract semantic and grammatical features by using basic semantic representation rules and setting different weights for words with different semantic contributions. In addition, the syntactic information is preserved by extracting word order features using convolutional operations on the context. The semantic representation of word vectors is further enhanced by using synonym, antonym, and analogy dictionaries, and corresponding correction strategies. In this paper, the effectiveness of the model is verified by similarity experiments, downstream entity recognition experiments, and ablation experiments. As a lightweight TCM text word representation model, the VCPC-WE model is as effective as mainstream deep pre-training models such as BERT. Considering the characteristics of TCM text, the next step is to consider formulating a series of omission completion and referential association rules to further enhance the performance of the model.

Author Contributions: Conceptualization, Y.M., D.Z. and Z.S.; methodology, Y.M. and Z.S.; software, Y.M., Z.S. and Y.F; validation, Y.M., Z.S. and Y.F.; investigation, Z.S.; resources, Y.M.; writing–original draft preparation, Z.S.; writing–review and editing, Y.M., Z.S. and Y.F.; supervision, Y.M.; project administration, Y.M., Z.S. and Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by Research on Common Technology and Method Technology System of Post-marketing Clinical Research of Traditional Chinese Medicine (Project-number: 2018YFC1707410), and Hebei Province 333 Talent Funding Project "Brain-like Intelligent Knowledge Discovery Technology Research" (Project-number: A201803082).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data for this study can be obtained by https://github.com/iwanttoknowmore/VCPC-WE-word-embeddings (accessed on 5 June 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
2. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 2014 Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 5–8 December 2013; pp. 3111–3119.
3. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
4. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
5. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
6. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
7. Lin, Y.; Tan, Y.; Frank, R. Open Sesame: Getting inside BERT's Linguistic Knowledge. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 1 August 2019; pp. 241–253.
8. Tenney, I.; Xia, P.; Chen, B. Wang, A.; Poliak, A.; McCoy, R.; Kim, N.; Das, B. What do you learn from context? Probing for sentence structure in contextualized word representations. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019; pp. 6–9.
9. Ettinger, T. What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 34–48. [CrossRef]
10. Zhang, C.; Lin, D.; Cao, D.; Li, S. Grammar guided embedding based Chinese long text sentiment classification. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e6439. [CrossRef]
11. Chang, Y.; Kong, L.; Jia, K.; Meng, Q. Chinese named entity recognition method based on BERT. In Proceedings of the 2021 IEEE International Conference on Data Science and Computer Application, Dalian, China, 29–31 October 2021; pp. 294–299.
12. Sun, M.; Yang, Q.; Wang, H.; Pasquine, M.; Hameed, I.A. Learning the Morphological and Syntactic Grammars for Named Entity Recognition. *Information* **2022**, *13*, 49. [CrossRef]
13. Wu, C.; Wu, F.; Liu, J.; Huang, Y.; Xie, X. Sentiment lexicon enhanced neural sentiment classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1091–1100.
14. Mao, X.; Chang, S.; Shi, J.; Li, F.; Shi, R. Sentiment-aware word embedding for emotion classification. *Appl. Sci.* **2019**, *9*, 1334. [CrossRef]
15. Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K.Q.; Artzi, Y. Revisiting Few-sample BERT Fine-tuning. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021; pp. 1484–1506.
16. Gao, Q. *Fundamentals of Unified Linguistics*; Science Press: Beijing, China, 2009.
17. Fan, X. *A Grammatical View of the Three Planes*; Beijing Language and Culture University Press: Beijing, China, 1996.
18. Fan, X. Research on verb–core structure. *Bull. Linguist. Stud.* **2011**, *1*, 1–34.
19. Jin, G. *Semantic Computation Theory of Verbs in Modern Chinese*; Peking University Press: Beijing, China, 2001.
20. Zhu, L.; Yu, T.; Yang, F. Study on semantic relations discovery based on key verbs in Chinese classical medical books. *China Digit. Med.* **2016**, *11*, 73–75.
21. Babiniotis, G. Towards a Linguistic Theory of Specification Based on a Verb Grammar. *Linguistics* **2022**, *10*, 176–180.
22. Qian, Z.; Lee, E.K.; Lu, D.H.Y.; Garnsey, S.M. Native and non-native (L1-Mandarin) speakers of English differ in online use of verb-based cues about sentence structure. *Biling. Lang. Cogn.* **2019**, *22*, 897–911. [CrossRef]
23. Zhou, H.; Hou, M.; Teng, Y. Construction research on opinion verbs-opinion targets intelligent computing. *J. Shanxi Univ. (Nat. Sci. Ed.)* **2022**, *45*, 274–283.
24. Liu, Q.; Ling, Z.; Jiang, H.; Hu, Y. Part-of-Speech Relevance Weights for Learning Word Embeddings. *arXiv* **2016**, arXiv:1301.3781.
25. Hu, B.; Tang, B.; Chen, Q.; Kang, L. A novel word embedding learning model using the dissociation between nouns and verbs. *Neurocomputing* **2016**, *171*, 1108–1117. [CrossRef]
26. Pan, B.; Yu, C.; Zhang, Q.; Xu, S.; Cao, S. The improved model for word2vec based on part of speech and word order. *Acta Electonica Sin.* **2018**, *46*, 1976–1982.
27. Wang, Z.; Liu, X.; Wang, L.; Qiao, Y.; Xie, X.; Fowlkes, C. Structured triplet learning with pos-tag guided attention for visual question answering. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1888–1896.
28. Deng, C.; Lai, G.; Deng, H. Improving word vector model with part-of-speech and dependency grammar information. *CAAI Trans. Intell. Technol.* **2020**, *5*, 276–282. [CrossRef]

29. Ren, X.; Zhang, L.; Ye, W.; Hua, H.; Zhang, S. Attention enhanced Chinese word embeddings. In Proceedings of the 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 154–165.
30. Yang, Y.; Mao, H.; Zhang, C. The impact of Inner Canon's language feature on its translation. *J. Zhejiang Bus. Technol. Inst.* **2015**, *14*, 80–88.
31. Qian, C. *Neijing Language Research*; People's Medical Publishing House: Beijing, China, 1990.
32. Wang, H. Study on the thought of numerology in *Huangdi Neijing*. Ph.D. Thesis, Beijing University of Chinese Medicine, Beijing, China, 2017.
33. Zhang, H. Discussion on the classification according to manifestation is the essence of Chinese medicine theory. *China J. Tradit. Chin. Med. Pharm.* **2016**, *31*, 4899–4901.