Modeling Multidimensional Language Matrices to Learn Predictive Text

Anonymous ACL submission

Abstract

The predictive text in the tray bed of the Chinese typewriter presents “Radiating style” and other important patterns, which reflect the main properties of the Chinese language. For a robot to understand these patterns like human’s “once glanced, never forgotten”, we construct multidimensional language matrices (MLM) to present the characters and/or words of predictive text for Chinese Natural Language Processing (NLP). Using 2D LM, our approach identified the core character as the prefix of radiating outward words, and as the suffix of radiating inward words to show the best distribution of the characters in a nine-grid. Using 3D LM, our approach, for robots doing as human, recognized the meaning and location of the words in a nine-grid by “Once learning mechanism”. Even though these approaches are proposed for the Chinese language, their methods are extendable to other languages.

1 Introduction

The predictive text relates the technologies to organize the characters in a panel or keypad to improve the input of messages or texts (MacKenzie, 2002; Tang et al., 2015; Sharma et al., 2019). The Chinese typewriter is a typical example of the development of predictive text (Lin, 1946; Mullaney, 2012). Especially in 1952, a typist reported some special patterns of “Radiating style and Connected thought” to implement the personalized distribution of Chinese characters in the tray bed of the Chinese typewriter (Li and Liu, 1952). With this arrangement, the typing speed reached 60 - 80 words per minute (WPM). Mullaney (2012; 2017) mentioned that it is possible to string the patterns of “Radiating style” in mini-regions together into ever-radiating associative networks to form a predictive text. Figure 1 shows *a)* a part of Simplified Chinese tray bed (SCTB) of the Chinese typewriter, in the 1960s-1990s, which is organized as predictive text; *b)* Actual Chinese character input panel in Google search, which still does not arrange as a predictive text. This picture shows that it has been over a hundred years, and the essence of Chinese typing/inputting has not changed by comparing the Chinese typewriter and the input panel of Google search (or Microsoft office, or smart mobile devices and facilities, etc.). The motivation of our research is to study the patterns of the predictive text of SCTB and transfer the technology to improve actual Chinese input systems.

Figure 1: *a)* Upper: a part of SCTB; *b)* Lower: the Chinese input panel for Google search

Even though some Intelligent Predictive Text Input Systems (IPTIS) are well developed for many languages (MacKenzie and Soukoreff, 2002; Elumeze and Nishimoto, 2006; Yang and Lin, 2014; Sim, 2014), these systems are developed for human using purposes. There are two kinds of problems or even challenges in the study of predictive text. The first is to arrange the characters and related words in the panel as the optimized predictive text. The second is to recognize the predictive text with “Radiating style” and other sophisticated patterns by computers or robots like humans do.

On the other hand, the technology of Chinese typewriting was a site of thoroughgoing, even with radical technolinguistic experimentation in the predictive text (Mullaney, 2012). There are few theoretical studies from the point of view of Natural Language Processing (NLP) in the literature.

In this research, firstly, we present the evolution of the tray bed of the Chinese typewriters from the Radical-Stroke arrangement to personalized predictive text; and then identify some kinds of patterns of Chinese predictive text, including the "Radiating style and Connected thought" using NLP methods. As the main contributions of our research, we construct multidimensional language matrices (MLM) to better present the Chinese language. Using MLM, we propose an approach to set the personalized predictive text by optimizing the distribution of characters in a nine-grid. We also developed a novel approach to identify Chinese words in "Radiating style" distribution by a nine-grid to simulate human's "Once learning mechanism (OLM)" behavior. OLM was proposed by (Weigang and da Silva, 1999) to model the "once seen, never forgotten(in Chinese: 过目不忘)". We present and prove some important properties and theorems of the fully connected nine-grid with "Radiating style". Even though our approaches are proposed based on the Chinese language, the novel methods are easily extendable to other languages.

2 Predictive Text in the Chinese Typewriter and Datasets

2.1 The Chinese typewriter

The Chinese typewriters have been available for more than a hundred years since 1912 (NYT, 1916; Zhang, 2008). The tray bed of the Chinese typewriter is designed with 2450 Chinese characters, alphabet letters, and numbers placed in 35 rows x 70 columns (SASS, 1983; Mullaney, 2012). Depending on the distribution of these characters, there are three generations of the tray bed.

1) The "Yu Bin-qi Chinese tray bed (YCTB)", 1920s-1940s: 9.8% of the spaces of the tray bed were arranged to put special and flexible characters (such as English letters), and the other parts followed the Radical-Stroke arrangement.

2) The "Traditional Chinese tray bed (TCTB)", 1950s: 11.18% of the spaces were arranged to put special characters, 16.65% of the spaces were used to mix the flexible, personalized with "Radiating style", and Radical-Stroke arranged characters in a common character region, and the other parts followed the Radical-Stroke arrangement.

3) The "Simplified Chinese tray bed (SCTB)", 1960s-1990s: 9.8% of the tray bed were arranged to put special characters; 37.47% were used to mix the flexible, personalized with "Radiating style and Connected thought", in two common character regions. The other parts followed the Radical-Stroke arrangement. SCTB is considered a predictive text. Table 1 shows some patterns of character distribution from SCTB.

Table 1: Some character patterns from the tray bed

No.	Example	Translation	Radiating Style	Pattern
1	年月日	Year Month Day	Horizontal	→→→
2	丁 丙 乙 甲	1,2,3,4 (Jia Yi Bing Ding – from down to up)	Vertical	↑ ↑ ↑ ↑
3	北西 南东	Southwest (西南) Northeast (东北)	Diagonal	↖↙ ↙↖
4	溶剂 液体	Solvent - horizontal Liquid - vertical	Azimuth	→ ↓
5	致般迫 些一切 唯统神	Identical Usually Press Some One All Only Unite	Nine-grids	↖↑↓ ← → ↗↑

"Radiating style" is a special pattern that sets characters in SCTB to improve the typing efficiency (Mullaney, 2012). For a better understanding of this style as a part of predictive text, we use 1-9 Chinese numbers to form a nine-grid panel combining 58 words by one character (e.g. 一^1, 二^2, ..., 九^9) or two characters (e.g. 二^{2}二^2, 三^{3}二^2, 二^{2}三^3, ...), see Figure 2. The superscript on the right side of the Chinese number/character is the translation. The words 五^{5}一^1, 五^{5}二^2, . . . , 五^{5}九^9 with 五^5 as prefix are combined by "Radiating outward". The words 四^{4}五^5, 七^{7}五^5, . . . , 八^{8}五^5 with 五^5 as suffix are combined by "Radiating inward"; and so on. The questions are: which number should be in the core and which ones in the other positions? How can a robot identify the nine-grids just by a glance (scanning)? We try to figure these problems in this research.

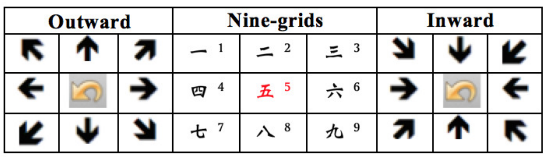

Figure 2: A *nine-grid* with "Radiating style" patterns

2.2 Datasets

The *modern Chinese corpus* was released by the State Language Commission of China (Xiao, 2012) and the online word search is also available.

1) Online Word Search from Corpus (OWSC). The corpus size is 20 million characters, where 162875 words are considered (151300 Chinese words).

2) Chinese Character Frequency Table (CCFT). The corpus size is 20 million characters, where 5708 characters with more than five occurrences are considered. The cumulative character frequency is 99.98%.

3) Chinese Word Frequency Table (CWFT). The corpus size is 20 million characters, where 14629 words with more than 50 occurrences are considered. The cumulative word frequency is 90.40%.

3 Language Vector and Matrix

Generally, there are two kinds of matrices to present natural language information: 1) the element is a letter/character or a word of the language; 2) the element is an index to describe the relationship between the letters/characters or words, such as mutual information (Cover and Thomas, 2006), word frequency (Xu et al., 2021), attention score (Seo et al., 2016; Vaswani et al., 2017), and others. Based on these researches, we first establish the Language Element Vector (LEV) and then define the Multidimensional Language Matrices (MLM).

An element refers to a basic component of a language $\mathcal{L}$ (Brill and Moore, 2000). In the case of alphabetical languages, a letter is an element, such as $a, b, c, ..., z, A, B, C, ..., Z$ of English. In the case of no alphabetical language, a character is an element, such as 一^1, 二^2, ..., 九^9 and so on.

Let $\sum$ be an alphabet of the language $\mathcal{L}$. Let $\sum^*$ be a set of all finite strings of the language $\mathcal{L}$. The dictionary D with the most commonly used words of $\mathcal{L}$ is considered as a subset of $\sum^*$ ($D \subseteq \sum^*$). The frequency of the word can be obtained from the corpus with a significant scale.

3.1 Definition of LEV

Definition 1 (LEV). The Language Element Vector (LEV) is defined as $\mathbf{V} = [z_1, z_2, z_3, ..., z_K]$, $z_k \in \sum$ is an element of a language $\mathcal{L}$, for $k = 1, 2, 3, ..., K$.

In the case of the nine-grid in Figure 2, z_k is a character from $\mathbf{V}_{nine}$ = [一^1, 二^2, ..., 九^9], $k = 1, 2, 3, ..., K, K = 9$.

In the case of the tray bed of the Chinese typewriter, $\mathbf{V}$ includes all characters in the tray bed, $K = 2450$.

As the Chinese language has the "general usage" property, $\mathbf{V}$ can include 3755 characters listed in the national standard GB2312-80, and the cumulative frequency of use is 99.7%. In this case, most Chinese texts can be expressed by these characters. This is an important advantage for us to develop multidimensional language matrices.

3.2 Definition of 2D LEM

Based on the Language Element Vector (LEV), we can construct various two-dimensional matrices depending on the purpose of the applications, such as Word Frequency Matrix and others.

Definition 2.1 (2D LEM). Two-dimensional language element matrix (2D LEM) is defined as, $M_{I\times J}$, with $w_{(i,j)}$-entry $\in \sum$, for $i = 1, 2, 3, ..., I$; and $j = 1, 2, 3, ..., J$;

In the case of the nine-grid in Figure 2, $w_{(i,j)}$ of $M_{3\times3}$, is [$w_{(1,1)} =$ 一^1, $w_{(1,2)} =$ 二^2, $w_{(1,3)} =$ 三^3, $w_{(2,1)} =$ 四^4, ..., $w_{(3,3)} =$ 九^9], $i = 1, 2, 3$; $j = 1, 2, 3$.

In the case of the tray bed of the Chinese typewriter, $M_{I\times J}$ includes all characters in the tray bed, for $i = 1, 2, 3, ..., I, I = 35$; $j = 1, 2, 3, ..., J$, $J = 70$; and $I \times J = K = 2450$.

Definition 2.2 (2D WFM). For a Language Element Vector, $\mathbf{V}$, if the characters z_k and z_{k1} in $\mathbf{V}$ combine a word $w_{(z_k, z_{k1})}$, k_1 is variant of k, the two dimensional word frequency matrix (2D WFM) is defined as $WFM_{K\times K}$, with $f(k, k_1)$-entry $\in$ $\mathbf{V^2}$, $f(k, k_1) = f_w(z_k, z_{k1})$, for $k = 1, 2, 3, ..., K$; and $k1 = 1, 2, 3, ..., K$. $f_w(z_k, z_{k1})$ is the frequency of the word $w_{(z_k, z_{k1})}$.

Definition 2.3 (2D WNM). Based on definition 2.2 (2D WFM), the two-dimensional word number matrix (2D WNM) is defined as matrix $WNM_{K\times K}$, with $a(k, k_1) = 1$, if $f(k, k_1) \geq \oslash$; $a(k, k_1) = 0$, if $f(k, k_1) < \oslash$; for $k = 1, 2, 3, ..., K$; k_1 is a variant of k and $k_1 = 1, 2, 3, ..., K$. $\oslash$ is a frequency threshold, where $0 \leq \oslash \leq 1$.

In the definition 2.2, the element $f(k, k_1)$ can also be the mutual information (Cover and Thomas, 2006), or the entropy of the word and others, depending on the necessity.

3.3 Application of 2D LEM

Figure 3 shows two nine-grids with Chinese characters by two patterns: 1) Radiating outward style, $\mathbf{V}_{out}$=[学^{learn}, 气^{air}, 会^{meet}, 小^{small}, 大^{large}, 约^{arrange}, 量^{quantity}, 概^{general}, 家^{family}]; 2) Radiating inward style, $\mathbf{V}_{in}$=[强^{powerful}, 伟^{great}, 广^{wide}, 增^{increase}, 大^{large}, 扩^{expand}, 长^{long}, 不^{no}, 重^{heavy}].

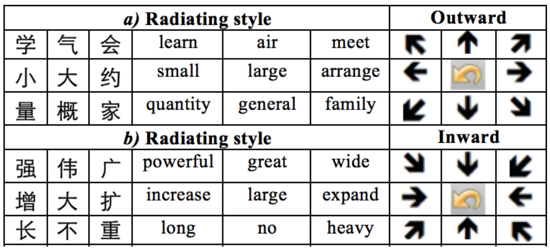

Figure 3: Two *nine-grids* with "Radiating style"

Using "Word Segmentation" by OWSC (Xiao, 2012), there are 12 words combined by two characters from Figure 3 a) Radiating outward. We present these words in 2D WNM as Table 2.

Table 2: Radiating outward in *nine-grid*

	学	气	会	小	大	约	量	概	家
学									
气									
会									
小	1						1		
大	1	1	1	1	1	1	1	1	1
约			1						
量									
概									
家									

Remark 1. For radiating outward case, in **2D WNM**, $a(k, k_1) = 1$, $k1 = 1, 2, 3, ..., 9$, refers to the word $w_{(z_k, z_{k1})}$ with z_k as a prefix.

In Table 2, the core character is "大^{large}" in $k = 5th$ row and almost all elements in this row with 1. Every one of these words related to this row has "大^{large}" as prefix: "大学^{university}", "大会^{meeting}", "大小^{LargeSmall}", "大大^{father}", etc.

Using "Word Segmentation" by OWSC (Xiao, 2012), there are 12 words combined by two characters from Figure 3 b) Radiating inward. We present these words in 2D WNM as Table 3.

Remark 2. For the radiating inward case, in **2D WNM**, $a(k, k_1) = 1$, for $k = 1, 2, 3, ..., 9$, refers to the word $w_{(z_k, z_{k1})}$ with z_{k1} as a suffix.

In Table 3, the core character is "大^{large}" in $k_1 = 5th$ column, and all elements in this column have value 1. Every one of these words in that column has "大^{large}" as suffix: "强大^{powerful}", "增大^{enlarged}", "广大^{vast}", etc.

3.4 Generalization 2D WNM for *Nine-grid*

The nine-grid is a basic unit to organize the characters in predictive text. We generalize 2D WNM and give more theoretical analysis.

Table 3: Radiating inward in nine-grid

	强	伟	广	增	大	扩	长	不	重
强					1				
伟					1				
广					1				
增	1				1		1		
大					1				
扩					1				
长					1				
不					1				
重					1				

In 2D WNM, two neighbor characters form a Chinese word, see Figure 3. We further identify three types of connections in a nine-grid: corner, tee, and core connections, as shown in Figure 4.

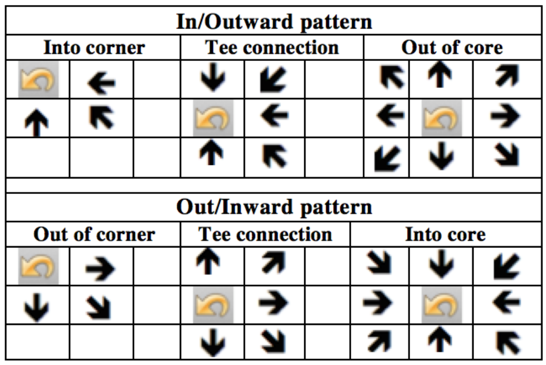

Figure 4: Three types of connections in *nine-grid*

1) In the core connection, the core character can connect to 8 other characters in two directions to form 16 words, plus itself as a single word (e.g., —^{1}) and the double of itself as a word (e.g., —^{1}—^{1}) there are 18 words in total.

2) In the tee connection, the character can connect to 5 neighbor characters, plus itself as a single word, itself doubled as a word, forming in 12 words (1 of them in common with core connection and 2 of them in common with other tee connections), and 32 words in subtotal.

3) In the corner connection, the corner character can connect to 3 neighbor characters, plus itself single as a word and itself doubled as a word, forming in 8 words (6 of them in common with others connections), and a subtotal of 8 words.

Definition 3 (Full connected nine-grid). A fully connected nine-grid consists of 49 words combined by two neighbor characters including the word by every character itself of 9. This nine-grid can be presented by a 2D WNM, see Table 6, which consists of 49 words combined by two neighbor characters including the word by repeating every character of 9. There are totally 58 possible words for a full connected nine-grid.

The 2D LEM of fully connected nine-grid is a symmetry matrix, but the most Chinese words related to this matrix are not symmetry. For example, “人家a kind of family” is different from “家人member of family”.

Table 4: 2D WNM of fully connected *nine-grid*

k/k_1	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	Sum
z_1	1	1		1	1					4
z_2	1	1	1	1	1	1				6
z_3		1	1		1	1				4
z_4	1	1		1	1		1	1		6
z_5	1	1	1	1	1	1	1	1	1	9
z_6		1	1		1	1		1	1	6
z_7				1	1		1	1		4
z_8				1	1	1	1	1	1	6
z_9					1	1		1	1	4
Sum	4	6	4	6	9	6	4	6	4	

Theorem 1 (Word connections in a nine-grid). Using 2D WNM to present fully connected nine-grid arranged by radiating style, the rank of three types of the word connections is listed as:

Rank 1. The core character as prefix connects to the other 8 characters separately, and these 8 words are listed in the $5th$ row; the core character as suffix connects to the other 8 characters, and these words are listed in the $5th$ column; the core character may repeat to form a word (same for tee and corner characters).

Rank 2. The tee character as prefix connects to the other 5 characters, these 6 words are listed in an even row, and there are four tee characters related to $2, 4, 6, 8th$ rows; the tee character as suffix connects to other 5 characters, these 6 words are listed in an even column, and there are four tee characters related to $2, 4, 6, 8th$ columns.

Rank 3. The corner character as prefix connects to the other 3 characters, these 4 words are listed in an odd row, and there are four corner characters related to $1, 3, 7, 9th$ row; the corner character as suffix connects to the other 3 characters, these 4 words are listed in an odd column, and there are four corner characters related to $1, 3, 7, 9th$ columns.

Proof. From 2D WNM of a fully connected nine-grid in Table 4, Rank 1:

$$\begin{aligned}\sum a(5,k_1) = max\Big(&\sum a(1,k_1), \sum a(2,k_1),\\ &..., \sum a(9,k_1)\Big)\\ =&(4,6,4,6,9,6,4,6,4) = 9,\end{aligned} \tag{1}$$

where, $a(5,k_1)$ relates to words with the core character z_5 as prefix, for $k_1 = 1, 2, ..., 9$.

$$\begin{aligned}\sum a(k,5) = max\Big(&\sum a(k,1), \sum a(k,2),\\ &..., \sum a(k,9)\Big)\\ =&(4,6,4,6,9,6,4,6,4) = 9,\end{aligned} \tag{2}$$

where, $a(k,5)$ relates to words with the core character z_5 as suffix, for $k = 1, 2, ..., 9$; $a(5,5)$ relates to the core character z_5.

We conclude that z_5 connects to neighbor characters to form a larger number of words than other arrangement.

Rank 2 and Rank 3:

From equation (1), we have,

$$\begin{aligned}&\sum a(2,k_1) = \sum a(4,k_1) =\\ &\sum a(6,k_1) = \sum a(8,k_1) = 6 >\\ &\sum a(1,k_1) = \sum a(3,k_1) =\\ &\sum a(5,k_1) = \sum a(7,k_1) = 4\end{aligned} \tag{3}$$

where, $a(2,k_1)$ relates to the words with the tee character z_2 as prefix; $a(1,k_1)$ relates to the words with the corner character z_1 as prefix; and so on, for $k_1 = 1, 2, ..., 9$.

From equation (2), we have,

$$\begin{aligned}&\sum a(k,2) = \sum a(k,4) =\\ &\sum a(k,6) = \sum a(k,8) = 6 >\\ &\sum a(k,1) = \sum a(k,3) =\\ &\sum a(k,7) = \sum a(k,9) = 4\end{aligned} \tag{4}$$

where, $a(k,2)$ relates to the words with the tee character z_2 as suffix; $a(k,1)$ relates to the words with the corner character z_1 as suffix; and so on; for $k_1 = 1, 2, ..., 9$.

In equations (3) and (4), we do not remove the repeated words but without loss of generality. When reducing the repeated words related to core, tee and corner connections, there are 49 words combined by two numbers/characters in the matrix. There are other possible 9 words formed by every single character in this *nine-grids*.

As 2D WNM in Table 4 is a general case for a fully connected *nine-grid*, we proved Theorem 1.

3.5 Definition of 3D LEM

2D LEM is proposed to present the words with two characters, such as "大会^{meeting}". For the words with three characters, such as "意大利^{Italy}", we can also span the vector V to three dimensions to get 3D LEM.

Definition 4 (3D LEM). Based on the Definition 1, the three dimensional LEV Span Matrices (3D LEM) is defined as matrix matrix $M_{K\times K\times K}$, with $b(k, k_1, k_2)$-entry $\in \mathbf{V}^3$, for $k = 1, 2, 3, ..., K$; k_1 and k_2 are variants of k and $k_1 = 1, 2, 3, ..., K$; $k_2 = 1, 2, 3, ..., K$. In **3D LEM**, $b(k, k_1, k_2) = 1$, if $f_w(z_k, z_{k_1}, z_{k_2}) \geq \oslash$; $b(k, k_1, k_2) = 0$, elsewhere; $f_w(z_k, z_{k_1}, z_{k_2})$ is the frequency of word $w(z_k, z_{k_1}, z_{k_2})$, $\oslash$ is a frequency threshold, $0 \leq \oslash \leq 1$.

4 Approach to set the nine-grids

In the nine-grid of Figure 2, which number should be in the core position? Generally, the number "一^1" has large frequency in the most of the languages. Can we put it in the core? In this section, we developed an approach to set the nine-grid using the numbers (1-9, in Chinese) in Figure 2 as an example.

Step 1. To construct a Word Frequency Matrix, $WFM_{9\times9}$, by V_{nine}=[一^1, 二^2, ..., 九^9] , with the element $f(k, k_1) = f_w(z_k, z_{k1})$, for $k_1 = 1, 2, ..., 9$.

Step 2. To determine the word frequency using the "Online word search from corpus" (Xiao, 2012). For example, the frequency of word "一^{1}一^1" in diagonal is 0.0024‰, "一^{1}二^2" is 0.0085‰, etc. Figure 5 shows the possible $WFM_{9\times9}$ (PWFM) by the characters from $\mathbf{V}_{nine}$ in a nine-grid.

k/k1	一 1	二 2	三 3	四 4	五 5	六 6	七 7	八 8	九 9	Sum
一 1	0.0024	0.0085	0.0004	0.0002	0.0019	0.0001	0.0002	0.0003	0.0004	0.0144
二 2	0.0007	0.0003	0.001	0.0003	0.001	0.0003	0.001	0.0014	0.0004	0.0064
三 3	0.0001	0.0002	0.0002	0.0022	0.0017	0.0002	0.0014	0.0052	0.0014	0.0126
四 4	0.0003	0.0006	0.0009	0.0003	0.0072	0.0006	0.0006	0.0004	0.0009	0.0118
五 5	0.0051	0.0005	0.0004	0.0151	0.0006	0.0024	0.0017	0.0015	0.0005	0.0278
六 6	0.0027	0.0005	0.0003	0.0006	0.0046	0.0005	0.0027	0.0009	0.0008	0.0136
七 7	0.0017	0.0008	0.0003	0.0007	0.0097	0.0004	0.0021	0.0026	0.0012	0.0195
八 8	0.0073	0.0013	0.0007	0.0018	0.0039	0.0005	0.0004	0.0003	0.0009	0.0171
九 9	0.0003	0.0004	0.0003	0	0.0009	0.0002	0.0003	0	0.0029	0.0053
Sum	0.0206	0.0131	0.0045	0.0212	0.0315	0.0052	0.0104	0.0126	0.0094	0.2570

Figure 5: $PWFM_{9\times9}$ by the characters in $\mathbf{V}_{nine}$(‰)

Step 3. To choose the character with high-frequency words as the core character of the nine-grid following Theorem 1. In this example, the sum of elements of $5th$ row is largest than other rows; this means that the character "五^5" as prefix connects to more characters. The sum of elements of $5th$ column is largest than other columns, which means that the character "五^5" as suffix also connects to more characters. So, in Chinese, the character "五^5" should be set as core in this nine-grid.

Step 4. To arrange other characters for the positions of the tee and corner. From $PWFM_{9\times9}$ in Figure 5, and the 2D $WNM_{9\times9}$ in Table 4, we can get the real $WFM_{9\times9}$ $(RWFM)$ by

$$RWFM = 2DWNM \times PWFM \tag{5}$$

Figure 6 shows the real WFM9x9 (PWFM) by the characters from $\mathbf{V}_{nine}$ in a nine-grid. The accumulated frequency of these 49 words is 0.1806‰, and the entropy of the words is 0.003978, where the frequencies of the repeated words are also calculated without loss of generality.

From the above 4 steps, we just determined the best core characters. This real WFM in equation 5 considers the neighbor position between two characters; it is not from the best arrangement of other characters in nine-grids.

k/k1	一 1	二 2	三 3	四 4	五 5	六 6	七 7	八 8	九 9	Sum
一 1	0.0024	0.0085		0.0002	0.0019					0.013
二 2	0.0007	0.0003	0.001	0.0003	0.001	0.0003				0.0036
三 3		0.0002	0.0002		0.0017	0.0002				0.0023
四 4	0.0003	0.0006		0.0003	0.0072		0.0006	0.0004		0.0094
五 5	0.0051	0.0005	0.0004	0.0151	0.0006	0.0024	0.0017	0.0015	0.0005	0.0278
六 6		0.0005	0.0003		0.0046	0.0005		0.0009	0.0008	0.0076
七 7				0.0007	0.0097		0.0021	0.0026		0.0151
八 8				0.0018	0.0039	0.0005	0.0004	0.0003	0.0009	0.0078
九 9				0	0.0009	0.0002		0	0.0029	0.004
Sum	0.0085	0.0106	0.0019	0.0184	0.0315	0.0041	0.0048	0.0057	0.0051	0.1812

Figure 6: $PWFM_{9\times9}$ by the characters in $\mathbf{V}_{nine}$(‰)

Step 5. Following Theorem 1, we adjust the distribution of the characters in tee and corner positions, see Figure 7 . The modified RWFM is with the accumulated word frequency 0.2216‰ and the entropy of the words 0,003247. Compared to Figures 6 , the accumulated word frequency increases 22.70%, and the entropy increases 22.53%. Figure 8 shows the modified nine-grids.

k/k1	二 2	一 1	六 6	八 8	五 5	七 7	四 4	三 3	九 9	Sum
二 2	0.0003	0.0007	0	0.0014	0.001	0	0	0	0	0.0034
一 1	0.0085	0.0024	0.0001	0.0003	0.0019	0.0002	0	0	0	0.0134
六 6	0	0.0027	0.0005	0	0.0046	0.0027	0	0	0	0.0105
八 8	0.0013	0.0073	0	0.0003	0.0039	0	0.0018	0.0007	0	0.0153
五 5	0.0005	0.0051	0.0024	0.0015	0.0006	0.0017	0.0151	0.0004	0.0005	0.0278
七 7	0	0.0017	0.0004	0	0.0097	0.0021	0	0.0003	0.0012	0.0154
四 4	0	0	0	0.0004	0.0072	0	0.0003	0.0009	0	0.0088
三 3	0	0	0	0.0052	0.0017	0.0014	0.0022	0.0002	0.0014	0.0121
九 9	0	0	0	0	0.0009	0.0003	0	0.0003	0.0029	0.0044
Sum	0.0106	0.0199	0.0034	0.0091	0.0315	0.0084	0.0194	0.0028	0.006	0.2222

Figure 7: Modified $RWFM_{9\times9}$

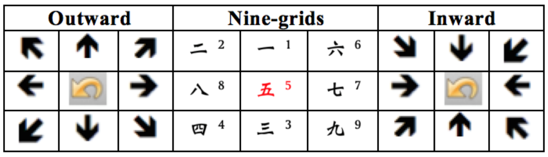

Figure 8: Modified nine-grids

There are 9! combinations of the distribution of 1-9 Chinese numbers in a nine-grid. We implemented an algorithm to analyze all 9! cases. As a result, there is Proposition 1.

Proposition 1 (Best distribution of a nine-grid). Considering the words formed by two Chinese numbers from "$一^1$, $二^2$, ..., $九^9$", the best distribution of these numbers in a nine-grid is with "$五^5$" as the core, "$一^1$,$三^3$,$七^7$,$八^8$" in tee, and others in corner positions separably. There are total 8 combinations, Figure 8 is one of them. In this case, the accumulated word frequency (or entropy) is maximum according to "Online word search from corpus" (Xiao, 2012).

5 Approach to recognize the nine-grids

In this section, we develop a new approach to identify the nine-grid in Figure 2 by robots.

5.1 Definition of 3D MLM

To understand the patterns of predictive text, we establish a Multidimensional language matrix (MLM) to present related characters, words, or frequencies of the words associated with the characters.

Taking the characters from nine-grid of Figure 3 a) as example, we first construct 2D $LEM_{3\times3}$ in $(X \times Y)$ plane as a reference, in X direction, $i = 1, 2, 3$; in Y direction, $j = 1, 2, 3$; $XY(1, 1) = z_1 =$ "$学^{learn}$" , $XY(1, 2) = z_2 =$ "$气^{air}$", ..., total with 9 characters. In Z direction, take vector $\mathbf{V}_{nine}[z_1, z_2, ..., z_9]$.

Definition 5 (3D LM). The 3D LM is defined as $Q_{3\times3\times9}$, with $q(i, j, k)$-entry, for $i = 1, 2, 3$; $j = 1, 2, 3$; $k = 1, 2, 3, ..., 9$. The element $q(i, j, k) = 1$ if $f_w(XY(i, j), z_k)e$; $q(i, j, k) = 0$, elsewhere. $\oslash$ is a frequency threshold, $0 \le \oslash \le 1$.

5.2 Recognize nine-grids using 3D LM

Following the "Once Learning Mechanism (OLM)" (Weigang and da Silva, 1999) to simulate this human behavior (过目不忘), we present an algorithm to recognize "Radiating style" of the Chinese language, such as nine-grids in Figure 3. Whoever knows the Chinese language will be able to "hunt" this pattern at first glance. The general image from this glance is a group of Chinese characters/words related to "$大^{large}$"as the core. Using nine-grids is only an example, and the algorithm can be extended to general cases.

[Step 1.] Take the characters from the nine-grid to form a text in the following 8 directions: left to right, right to left, up to down, down to up, up left to down right, down right to up left, up right to down left, down left to up left, and repeating itself for every character. If using parallel processing, the reading in 8 directions can be processed at the same time. In the Figure 3 , we get the Text.Out from *a)* Outward and the Text.In from *b)* Inward.

[Step 2.] Use a function of "Word Segmentation" of NLP to separate the words. By OWSC (Xiao, 2012), we obtained 12 words from Text.Out, see Table 2, and 12 words from Text.In, see Table 3.

[Step 3.1] Following the Definition 5, take the 9 characters from nine-grid in Figure 3 a) Outward style: $\mathbf{V}_{out}$ =[$学^{learn}$, $气^{air}$, $会^{meet}$, $小^{small}$, $大^{large}$, $约^{arrange}$, $量^{quantity}$, $概^{general}$, $家^{family}$] to form the 3D LM, $TEXTO_{3\times3\times9}$. The elements of this matrix are similar to Table 2, which is the reference to form an $(X \times Y)_{3\times3}$ plane. The Z-axis is formed by the elements of $\mathbf{V}_{out}$.

[Step 3.2] Also, following the Definition 5, take the 9 characters from nine-grids in Figure 3 b) Inward style: $\mathbf{V}_{in}$=[$强^{powerful}$, $伟^{great}$, $广^{wide}$, $增^{increase}$, $大^{large}$, $扩^{expand}$, $长^{long}$, $不^{no}$, $重^{heavy}$], to form the 3D LM, $TEXTI_{3\times3\times9}$. The elements of this matrix are similar to Table 3, which is a reference to form an $(X \times Y)_{3\times3}$ plane. The Z-axis is formed by the elements of $\mathbf{V}_{in}$.

[Step 4.1] Identify the pattern in 3D LM, $TEXTO_{3\times3\times9}$. In Radiating outward style, $q(2, 2, k) = 1$, for $k = 1, 2, ..., 9$, see Figure 9 a). In this case, a machine can identify the core character "大large" as prefix connecting to other 8 characters, such as "$大学^{university}$" $q(2, 2, 1) = 1$, "$大气^{air}$", "$q(2, 2, 2) = 1$", "$大大^{father}$ $q(2, 2, 5) = 1$", and others. There is $\sum q(2, 2, k) = 9$, for $k = 1, 2, 3, ..., 9$. There are "$小学^{school}$" $q(2, 1, 1) = 1$, "$小量^{Small\ amount}$" $q(2, 1, 7) = 1$, "$约会^{Dating}$" $q(2, 3, 3) = 1$ and other $q(i, j, k) = 0$.

[Step 4.2] Identify the pattern by 3D LM, $TEXTI_{3\times3\times9}$. In Radiating inward style, $q(i, j, 5) = 1$, for $i = 1, 2, 3$ and $j = 1, 2, 3$, see Figure 9 b). In this case, a robot can identify the core character "$大^{large}$" as suffix connect-

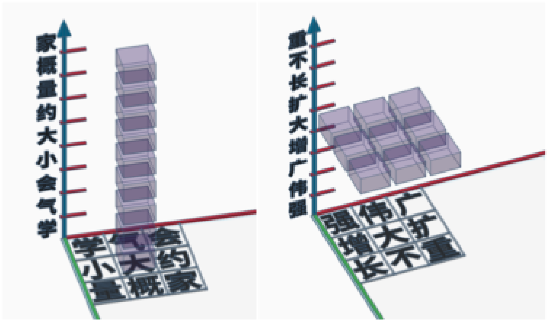

Figure 9: a) Left - radiating outward; b) Right - radiating inward in 3D LM

ing to other 8 characters, such as "强大^{powerful}" $q(1,1,5) = 1$, "伟大^{great}" $q(1,2,5) = 1$, "增大^{increase}" $q(2,1,5) = 1$, "大大^{father} $q(2,2,5) = 1$", and others. There is $\sum q(i,j,5) = 9$, for $i = 1,2,3$ and $j = 1,2,3$. There are "增强^{strengthen}" $q(2,1,1) = 1$, "增长^{grow}" $q(2,2,7) = 1$, and other $q(i,j,k) = 0$.

[Step 5.] With the results from the algorithm, the robots can identify the characters and their positions in the nine-grid with both radiating outward and inward styles.

The above example is only for one nine-grid. We can apply the approach to more nine-grids in predictive text. Figure 10 presents two nine-grids with both radiating outward and inward styles. For better understanding, we put letters $[A, B, ..., R]$ as an example. In both cases, the core letters can be easily identified, in this case, H and K. The application of our approach can help computers to better understand the radiating style and the relationship between the letters/characters to form the words for a language.

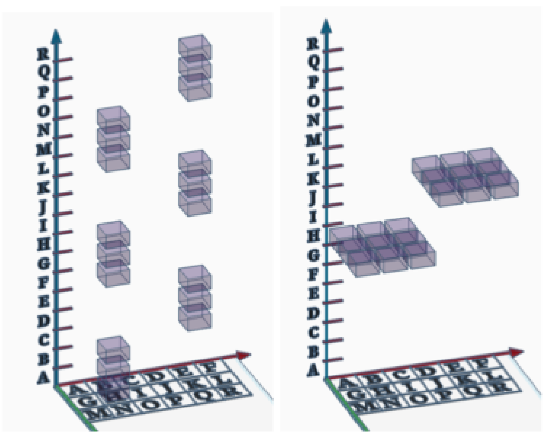

Figure 10: Two nine-grids: a) Left - radiating outward; b) Right - radiating inward in 3D LM

6 Conclusions

The unit of nine-grids is the basic pattern in predictive text with "Radiating style", especially for the Chinese language. We identify the word patterns in nine-grids including: horizontal, vertical, and diagonal, in double directions to form the Radiating outwards and inwards. To better present these patterns, the multidimensional language matrices (MLM) are established based on the previous literature.

From the 2D LEM, we demonstrated that the core character as prefix connects to neighbor characters by radiating outward in nine-grid, and also the core character as suffix connects to neighbor characters by radiating inward in nine-grid.

After identifying three connection types in a nine-grid, we proved the rank of the connectivity among core, tee, and corner positions. Based on this theorem, we construct a fully-connected matrix for a nine-grid with the maximum number of 58 words formed by one or two characters and propose the approach to set the nine-grids with the best character combination.

We also proved the best combination of Chinese numbers [一^1, 二^2, ... , 九^9] in a nine-grid as the matrix in Figure 7 and Figure 8. Especially, the "五^5" should be in the core position with maximum accumulated word frequency (or entropy) being both prefix and suffix to other numbers.

It is important to simulate the human's behavior to see the predictive text and extract the information by a computer. We proposed an approach to recognize the nine-grids by"Once learning mechanism". The developed algorithm can identify the core character and the connected characters in one or more nine-grids. The manner of the identification is demonstrated by 3D space in Figure 9 with one nine-grids and Figure 10 with two nine-grids for both radiating outward and inward. To the best of our knowledge, this presentation is novel.

The research is based on the predictive text from the Chinese typewriter. In the search panel for "Google", in Figure 1 b), there are 6x6 grids to present one or two characters. The approaches we developed can be applied to set the panel in "Radiating style" or other optimized patterns for future research. Even though the developed approaches are proposed for Chinese language processing, their methods are easily extendable to other languages.

References

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. pages 286–293.

Thomas M Cover and Joy A Thomas. 2006. Elements of information theory second edition solutions to problems. *Internet Access*, pages 19–20.

Nwanua Elumeze and Keisuke Nishimoto. 2006. Intelligent predictive text input system using japanese language. *Final Report for CSCI 5832: Natural Language Processing*.

Zhongyuan Li and Zhaolan Liu. 1952. Kaifeng typesetter zhang jiying's advanced work method. *People's Daily*.

Yu Tang Lin. 1946. Invention of a chinese typewriter. *Asia and the Americas*, 46(4):58–61.

I Scott MacKenzie. 2002. Kspc (keystrokes per character) as a characteristic of text entry techniques. pages 195–210.

I. Scott MacKenzie and R. William Soukoreff. 2002. Text entry for mobile computing: Models and methods, theory and practice. *Human–Computer Interaction*, 17(2-3):147–198.

Thomas S. Mullaney. 2012. The moveable typewriter: How chinese typists developed predictive text during the height of maoism. *Technology and Culture*, 53(4):777–814.

Thomas S Mullaney. 2017. *The Chinese typewriter: A history*. MIT Press.

NYT. 1916. Chinaman invents chinese typewriter using 4,000 characters. *The New York Times*.

SASS. 1983. Double pigeon brand chinese typewriter - shanghai economics: 1949-1982. *Editorial Office of "Shanghai Economic Yearbook"*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *In the proceedings of ICLR 2017*.

Radhika Sharma, Nishtha Goel, Nishita Aggarwal, Prajyot Kaur, and Chandra Prakash. 2019. Next word prediction in hindi using deep learning techniques. pages 55–60.

Khe Chai Sim. 2014. A multimodal stroke-based predictive input for efficient chinese text entry on mobile devices. pages 448–453.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. pages 1165–1174.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.

Li Weigang and Nilton Correia da Silva. 1999. A study of parallel neural networks. In *IJCNN'99. Int Joint Conf on Neural Networks.*, volume 2, pages 1113–1116.

Hang Xiao. 2012. Introduction to the modern chinese corpus of the state language commission (国家语委现代汉语语料库介绍). *Accessible time: October 5, 2021. http://corpus.zhonghuayuwen.org/index.aspx*.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373.

Heng-Li Yang and Ren-Xiang Lin. 2014. A chinese predictive text entry method for mobile devices. 1:309–314.

Yao Jun Zhang. 2008. Shu zhendong and china's first chinese typewriter (舒震东与中国第一台中文打字机). *Shanghai Archives Information Network*.

A Appendix: Text.In and Results from Word Segmentation

A.1 Text.In (in Chinese)

Take the characters from the nine-grid to form a text in the following 8 directions: left to right, right to left, up to down, down to up, up left to down right, down right to up left, up right to down left, down left to up left, and repeating itself for every character. If using parallel processing, the reading in 8 directions can be processed in parallel. Figure 3, we get the Text.In from b) Inward.

强伟广增大扩长不重广伟强扩大增重不长强增长伟大不广扩重长增强不大伟重扩广强大重广大长重大强长大广强强伟伟广广增增大大扩扩长长不不重重

A.2 Results from Online Word Segmentation from Text.In

Using "Word Segmentation" function by Online Word Search from Corpus (Xiao, 2012), (corpus.zhonghuayuwen.org/CpsTongji.aspx), we get 12 words combined by two characters, see Table 5.

B Appendix: The pseudo and python code of algorithm

Table 5

No.	Word	Translation	Occurrence	Frequency %
1	广	wide	8	12.6984
2	强	powerful	7	11.1111
3	长	long	7	11.1111
4	不	no	6	9.5238
5	扩	expand	6	9.5238
6	伟	great	6	9.5238
7	重	heavy	6	9.5238
8	增	increase	4	6.3492
9	不大	not large	1	1.5873
10	大	large	1	1.5873
11	大大	father	1	1.5873
12	广大	vast	1	1.5873
13	扩大	expand	1	1.5873
14	强大	powerful	1	1.5873
15	伟大	great	1	1.5873
16	增大	increase	1	1.5873
17	增强	enhance	1	1.5873
18	增长	increase	1	1.5873
19	长大	grow up	1	1.5873
20	重大	major	1	1.5873
21	重重	numerous	1	1.5873