CIT. Journal of Computing and Information Technology, Vol. 29, No. 2, June 2021, 113–131
doi: 10.20532/cit.2021.1005393

Natural Language Processing Using Neighbour Entropy-based Segmentation

Jianfeng Qiao[1, 2], Xingzhi Yan[3] and Shuran Lv[1]

[1]Capital University of Economics and Business, Beijing, China
[2]Beijing Key Laboratory of Megaregions Sustainable Development Modeling, Beijing, China
[3]University of Birmingham, UK

In natural language processing (NLP) of Chinese hazard text collected in the process of hazard identification, Chinese word segmentation (CWS) is the first step to extracting meaningful information from such semi-structured Chinese texts. This paper proposes a new neighbor entropy-based segmentation (NES) model for CWS. The model considers the segmentation benefits of neighbor entropies, adopting the concept of "neighbor" in optimization research. It is defined by the benefit ratio of text segmentation, including benefits and losses of combining the segmentation unit with more information than other popular statistical models. In the experiments performed, together with the maximum-based segmentation algorithm, the NES model achieves a 99.3% precision, 98.7% recall, and 99.0% f-measure for text segmentation; these performances are higher than those of existing tools based on other seven popular statistical models. Results show that the NES model is a valid CWS, especially for text segmentation requirements necessitating longer-sized characters. The text corpus used comes from the Beijing Municipal Administration of Work Safety, which was recorded in the fourth quarter of 2018.

ACM CCS (2012) Classification: Information systems → Information retrieval → Retrieval models and ranking → Language models

Keywords: text mining, text segmentation, Chinese word segmentation, safety management, hazard analysis

1. Introduction

In order to prevent accidents in workplaces, there is usually a mechanism in place for workers to report hazards before they occur [1]. Through rectifying the hazards, the risks of similar events can be mitigated [2]. In January 2017, the General Office of the State Council of China issued the 13th Five-Year Plan for Safety Management, which emphasized hazard identification as the most critical task for safety management in China. *Hazard texts* (or *hazard reports*) recorded by safety management workers are the essential sources of hazard identification. However, the amount of hazard texts that need to be read and analyzed in Chinese is so large that they are beyond the analytical capability of humans, hence tools or system should be employed to help the respective analysis. Since it has been noted that the efficiency of existing tools is poor, a novel approach is proposed and elaborated in the continuation of the paper.

1.1. Hazard Text

Hazard reports are recorded during the hazard identification process according to specific attributes or structures. Each record includes the checking time, the name of the workshop being investigated, the content of the hazard item (*i.e.* the hazard text), and the category of hazard, as shown in Table 1.

A hazard report has the following characteristics:

1. the report itself is semi-structured, which on the one hand means that it is saved in a

structured storage according to the checking time and other parameters; while on the other hand, the hazard text is narrative (*i.e.* un-structured), and the length of each segment is uncertain;

2. the hazard text is a kind of short text, and each of such texts can be further divided into a shorter sentence by using a sequence number, and
3. the hazard text contains professional words, generally described in terms of safety science, such as "安全生产" (work safety) and "责任制度" (responsibility system).

1.2. Hazard Text Segmentation

Safety management information systems typically have to process quite an amount of textual data like the hazard text. Fortunately, with advances in data mining techniques, there has been a continuous growth in text mining knowledge in the injury field [3].

Currently, the majority of studies focus on information retrieval (IR), natural language processing (NLP) [4–7], and text classification [8–11], but there is only limited research on text segmentation, which can be further subdivided into word segmentation, feature segmentation, and sentence segmentation. Text segmentation is an important NLP task, which is fundamental to many text mining tasks, here ncluding text classification and text clustering. It also has an essential effect on IR effectiveness [12]. Text segmentation results facilitate detailed analysis by automatically extracting knowledge.

Text objects are generally derived more from compensation claims, accident reports, and near-miss reports and less from hazard texts. However, like other unstructured or semi-structured texts, the analysis of hazard texts can help extract more information in decision-making and support safety management [13, 14] to better serve accident analysis and prevention.

There are several well-known CWS tools; however, the respective segmentation results are not optimized for hazard texts, the primary reason stemming form the fact that most segmented units are composed of unigram and bigram characters. As a matter of fact, in modern standard Chinese there are 5% one-character words, 75% two-character words, and 14% three-character words, with 6% being four- or more characters [15]. These cannot meet the hazard text segmentation requirements, requiring longer-sized characters. For the Chinese, text segmentation is the primary step for many downstream tasks, and some suitable methodologies with high performance and low computational cost need to be investigated systematically.

For Chinese text mining, a fundamental task is to perform Chinese word segmentation (CWS), which is also a critical process in the initial stages of other Chinese text mining approaches. Text segmentation, which partitions a plain text into a sequence of meaningful words, is one of the most critical research areas in text

Table 1. An example of a record of hazard reports.
(above: original Chinese text; below: English translation).

Checking time	Name of workshops	Hazard texts	Categories
2018-10-25	北京XX餐饮有限公司	1、未健全安全生产责任制度; 2、从业人员董雨倩等38人未经安全生产教育上岗工作;	1. 安全生产责任制缺陷 2.培训教育不足
10-25-2018	Beijing XX Catering Co., Ltd.	1. Need to improve the responsibility system for work safety. 2. A total of 38 employees work without safety trainings, such as Dong Yuqian	1. Defects of responsibility systems 2. Without enough training

Note: "安全生产" is a compact unit with four characters (long characters); "董雨倩" is a person's name (out-of-vocabulary word).

mining. In English and many other languages that use some form of Latin alphabet, the space is a good word delimiter. However, some Asian texts only consist of character strings without obvious boundaries between words, except for punctuation signs at the end of each sentence and occasional commas in the sentences. In addition, the ambiguous and out-of-vocabulary (OOV) words make it difficult to accurately separate a string of Asian characters into words. Ambiguous words are composed of one or more characters and are not explicitly delimited [16].

There are at least two special characteristics of Chinese hazard text segmentation. First, the length of word segmentation is considerable and dynamic, such as "未健全" (3 characters) and "安全生产" (4 characters). Generally, the length of word segmentation contains more unigram and bigram characters in other fields. Second, it shows the problems of ambiguous segmentation and of named entity recognition (one of the OOV problems), such as a person's name ("董雨倩", Dong Yuqian) and place name, as displayed in Table 1.

1.3. General Methods of CWS

There are six types of segmentation methods: dictionary-based, statistics-based, intelligent algorithm-based, in-depth learning-based, hybrid-based, and tool-based methods.

1. The *dictionary-based* word segmentation methods, such as maximum forward and reverse matching, have a poor recognition rate for unfamiliar words and ambiguities. Although the speed of segmentation is fast, the accuracy of matching methods is limited [17].
2. The *statistical model* is the most straightforward and efficient method. N-gram probability and mutual information (MI) are the typical models [18, 19]. Bigram and MI are measures of co-occurrence between characters, where the fewer the co-occurrence between two continuous characters, the larger the probability of separating them. Therefore, these models have limitations when handling the OOV words (sometimes occurring only once) and ambiguous words (sometimes combined with the lower probability).
3. Many *neural network-based* methods have been proposed for CWS [20]. It is one kind of distributed *n*-grams. The dynamic programming algorithm is applied, and it requires storing lots of information about previous steps. The maximal length of the segments affects the performance, which makes it somewhat complex compared with simple statistic methods. Particle swarm optimization (PSO) is another popular intelligent algorithm in text segmentation [21]. However, PSO fitness functions are difficult to establish for the benefit of text segmentation, and applying the algorithm to more significant dimensional optimization problems is also problematic [22].
4. Many *in-depth learning-based* methods have also been put forward for text segmentation [23, 24], but they require many labeled sentences during model training. Some semi-supervised deep learning methods have also been built [25]. However, compared to shallow learning methods, they are black-box models, complicated corresponding algorithms, and expensive computations in training and prediction.
5. In addition, *hybrid approaches* usually combine two approaches to merge the benefits of general and domain-specific knowledge [26]. Although the hybrid approaches take advantage of different approaches to obtain more accurate segments, these are achieved at the expense of increasingly complex processing times, disk spaces, and cost requirements. The balance between capacity and computational cost is challenging for using most machine learning approaches.
6. Certainly, the simplest method is to *adopt available CWS tools*, such as ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). However, segmentation results provided by these popular SW tools are not applicable for hazard text because most auto-segmented units are composed of unigram and bigram characters, which are not the information we seek to extract. While the hazard text is more likely to be segmented as more extended size characters, it also implies

that the segmented unit should contain a trigram, 4-gram, or more characters, because they are compact and meaningful units. For example, "安全生产" (work safety) is a compact and meaningful unit, so it does not need to be segmented as "安全|生产." In addition, these CWS tools use dictionary-based techniques, and some OOV words from the injury field are not included. Moreover, some words in the dictionary are not suitable for the injury field, thus leading to lower accuracy of text segmentation in the injury field.

1.4. Improved Understanding from Optimization

Switching to a new field with many new words beyond the considered dictionary makes the dictionary-based word segmentation methods less suitable. The statistical model is still the priority when baselining segmentation results quickly.

Currently, most statistical models only focus on their combined benefits and do not consider losses when combined with other data particles (or compact data units, such as Chinese characters). For example, the benefit of an OOV word is lower because of its low probability. However, if the characters in the OOV words are combined with their neighboring characters, it could cause the loss of the whole text stream or lessen the benefit for the whole text stream during the segmentation process. It means that focusing on the fitness value of the unit is not enough during the text segmentation process, and we have to consider the benefits of neighbors.

In optimization research, the fitness function is established by itself and its "neighbors" [27]. Several text segmentation models are composed of information entropy. Therefore, this paper proposes a *neighbor entropy-based segmentation* (NES) to measure the benefit and loss of different segmentation choices.

A simple maximum-based segmentation algorithm is designed for CWS to realize the proposed model. The proposed solution is compared with the CWS tool and other popular statistical models by applying a test corpus taken from the collected texts during the hazard identification process for some catering enterprises in Beijing. Further to text segmentation, some helpful information is retrieved. It can be seen that there are several common outstanding hazards for catering enterprises, which will provide help and support for making decision and safety management.

Hazard analysis is a crucial task in accident analysis and prevention. Automatic text segmentation represents a fundamental task in text mining to extract meaningful information from semi-structured hazard text, where meaningful information means the valid feature word segmented by text segmentation. The high dimensional space causes the problem of text visualization. After the segmentation, the hazard text is easier to read and understand, thus allowing safety managers to focus on prevention strategy. Moreover, text segmentation is necessary to reduce the feature dimension required by text classification and clustering in NLP. Due to current CWS tools' limitations, we have to develop new text segmentation methods suitable for the injury field. The statistical method prioritizes its simplicity and effectiveness when switching to a new field.

Furthermore, such method is a baseline method for performance comparison. However, the existing state-of-the-art statistical methods still have limitations, as mentioned above. Hence, we need to investigate the improvement based on theoretical limitations of the models themselves and the characteristics of texts in the injury field.

2. Related Work

2.1. Concepts

Before discussions, let us clarify two concepts: hazard text and text segmentation. Many organizations have significant sources of textual data, including hazard texts, near-miss reports, accident reports (or injury reports), and accident investigation reports. *Hazard texts* provide information about risks; they provide a valuable complement to an accident reporting system. The near-miss report identifies various hazards within an organization or industry. The coding of near misses will help construct hazard sce-

narios and the informed development of appropriate interventions to prevent future injuries. The near-miss report is a narrative text, while the hazard text is semi-structured, and composed of short sentences. Both near-miss and accident reports have great significance in surveillance systems. While injury data gives information about incidence and the direct cause of an injury, near-miss data enable the identification of various hazards within an organization or industry while providing an opportunity for risk reduction. In simple terms, "yes" means injury, and "no" means near-miss [8].

Text segmentation provides a solution for extracting meaningful units from unstructured texts. These compact sets of units (or particles) can be words, features, or subtopics. According to the categories of compact units, text segmentation can be divided into word segmentation, feature segmentation, and sentence segmentation. For *word segmentation* (text segmentation means word segmentation in this paper), the segmentation unit is a word composed of several characters. Thus, word segmentation aims to segment plain text into a sequence of meaningful words [28]. The word concept is a sequence of characters delimited by precise word boundaries [29]. *Feature segmentation* is the problem of dividing written words into key phrases (two or more groups of words). This process converts a sentence into meaning features, terms, or tokens; the segmentation result improving the performance of text classification and clustering [30]. *Sentence segmentation* is splitting a long document or a text stream into several topical segments, usually blocks of sentences. In other words, sentence segmentation aims to identify topic boundaries within a document to partition it into several topical segments, each consisting of consecutive sentences [31]. It is applied in the first step of document summarization, which often breaks a document into topics.

2.2. Statistical Segmentation Models

Chinese word segmentation is the first step for extracting valuable information from the original (Chinese) hazard text, serving as the basis for applying subsequent text mining techniques. Several efficient statistical language models segment unpartitioned texts into coherent fragments.

Model 1, one of the most popular models for the NLP, is the *n*-gram probability. *N*-gram refers to a set of *N* adjacent elements as they appear in texts. Multi-character-based approaches segment texts into strings containing one (unigram), two (bigram), three (trigram), or more (*n*-gram) characters. Given two continuous random variables w_i and w_{i+1}, their bigram probability $p(w_i, w_{i+1})$ is defined in terms of probabilistic density functions and conditional probability:

$$p(w_i, w_{i+1}) = p(w_i)p(w_{i+1} \mid w_i) \tag{1}$$

where $p(w_{i+1} \mid w_i) = \dfrac{p(w_i, w_{i+1})}{p(w_i)}$.

Using the maximum likelihood estimation (MLE) method,

$$p(w_i) = \frac{c(w_i)}{\text{TotalNum}};\ p(w_i, w_{i+1}) = \frac{c(w_i, w_{i+1})}{\text{TotalNum}}.$$

TotalNum is the total number of variables of the text; $c(w_i)$ and $c(w_i, w_{i+1})$ are the count number of the arguments (such as Chinese characters). The conditional probability $p(w_i, w_{i+1}) \leq 1$. If $c(w_i, w_{i+1})$ is almost equal to $c(w_i)$, the conditional probability has the highest value. The higher the value of bigram probability, the greater the possibility of it being a compact unit.

For example, if "安" (a Chinese character) occurs 20 times, and "安全" (safety) occurs ten times in the corpus, then

$$p(\text{安, 全}) = p(\text{安})p(\text{全} \mid \text{安}) = \frac{20}{\text{TotalNum}} \cdot \frac{10}{20}.$$

Model 2 is the *t*-model based on *t*-distribution [32]. It considers the contextual relationship of w_i:

$$t(w_i) = \frac{p(w_{i+1} \mid w_i) - p(w_i \mid w_{i-1})}{\sqrt{\delta^2\left(p(w_{i+1} \mid w_i)\right) + \delta^2\left(p(w_i \mid w_{i-1})\right)}} \tag{2}$$

where variance $\delta^2\left(p(w_{i+1} \mid w_i)\right) = \dfrac{c(w_i, w_{i+1})}{c^2(w_i)}$

and $\delta^2\left(p(w_i \mid w_{i-1})\right) = \dfrac{c(w_{i-1}, w_i)}{c^2(w_{i-1})}$.

Model 3, another type of statistical model that breaks the input text stream into candidate words according to entropy information, is based on word boundary entropy and is denoted by H_B [33]:

$$H_B(w_i) = -\sum_{w_{i+1} \in C} p(w_{i+1} \mid w_i) \ln\left(p(w_{i+1} \mid w_i)\right) \quad (3)$$

where C is the set of all possible successor characters following a subsequence w_i, and $p(w_{i+1}) \mid w_i)$ is the occurrence probability of character w_{i+1} following w_i.

Model 4 is based on MI, a measure of co-occurrence between characters [19]:

$$I(w_i, w_{i+1}) = p(w_i, w_{i+1}) \log \frac{p(w_i, w_{i+1})}{p(w_i)p(w_{i+1})} \quad (4)$$

The MI definition can be rewritten in terms of entropies and conditional entropies as follows:

$$\begin{aligned} I(w_i, w_{i+1}) &= H(w_i) - H(w_i \mid w_{i+1}) \\ &= H(w_{i+1}) - H(w_{i+1} \mid w_i) \end{aligned} \quad (5)$$

It can take values in the following interval:

$$0 \le I(w_i, w_{i+1}) \le \min\left\{H(w_i), H(w_{i+1})\right\} \quad (6)$$

Then, model 5 is defined as normalized MI [34]:

$$I_N(w_i, w_{i+1}) = \frac{I(w_i, w_{i+1})}{\min\left\{H(w_i), H(w_{i+1})\right\}} \quad (7)$$

Model 6 is defined as symmetric MI to measure symmetric uncertainty [35]:

$$I_S(w_i, w_{i+1}) = \frac{2I(w_i, w_{i+1})}{H(w_i) + H(w_{i+1})} \quad (8)$$

I_N and I_S take values between 0 and 1, respectively, where 0 indicates that the variables are independent and '1' implies that the knowledge about the value of one variable is enough to define the other perfectly.

In statistics and probability theory, correlation measures dependence between two paired random vectors. The correlation concept has been used in text-similarity measurement [36] and features dependency analysis [37]. Since the similarity between data particles is often regarded as a symmetrical relationship, model 7 can be defined as data correlation-based segmentation (DCS):

$$\begin{aligned} Cor(w_i, w_{i+1}) &= p(w_{i+1} \mid w_i)p(w_i \mid w_{i+1}) \\ &= \frac{p(w_i, w_{i+1})}{p(w_i)} \frac{p(w_i, w_{i+1})}{p(w_{i+1})} \end{aligned} \quad (9)$$

$$Cor(w_i, w_{i+1}) = \frac{c(w_i, w_{i+1})^2}{c(w_i)c(w_{i+1})} \quad (10)$$

In summary, apart from popular models such as n-gram probability in Equation (1) and MI in Equation (4), the above section also lists five additional statistical models: *t*-model in Equation (2), word boundary entropy in Equation (3), normalized MI in Equation (7), symmetric MI in Equation (8), and DCS in Equation (9); indeed, the fitness values of the above models are calculated by these equations separately. All seven models achieve good performances for their own NLPs.

Word segmentation must rely on character-level information indicated by conditional probability, such as *n*-gram probability, or co-occurrence probability, such as MI. MI can also be written as:

$$\begin{aligned} I(w_i, w_{i+1}) &= p(w_i)p(w_{i+1} \mid w_i) \\ &\log \frac{p(w_{i+1} \mid w_i)}{p(w_i)} \end{aligned} \quad (11)$$

The higher the conditional probability is, the more significant the MI. Both models, *n*-gram probability and MI, only consider conditional probability $p(w_{i+1} \mid w_i)$, which shows whether w_i would like to combine it with w_{i+1}. It is not considered whether w_{i+1} would like to combine it with w_i.

T-model and word boundary entropy models are also built from conditional probability while considering contextual relationships; this means that when setting the word boundary for the current character, they compare with the following variable and the former variable.

Normalized MI, symmetric MI, and DCS are defined symmetrically. For any two continuous particles, there is a backward conditional

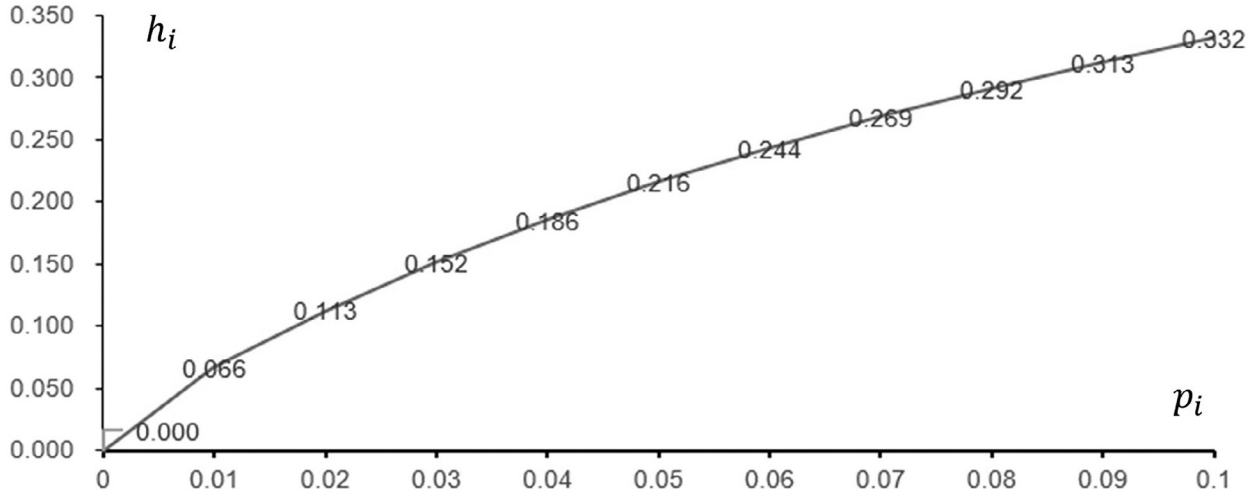

Figure 1. Entropy value curve.

probability $p(w_{i+1} \mid w_i)$ for w_i and a forward conditional probability $(w_i \mid w_{i+1})$ for w_{i+1}. If conditional probabilities of both directions are high, particles should be combined with strong correlations. Compared with bigram probability and MI, the symmetric models consider both directions. Theoretically, they will be more efficient than n-gram probability and MI, which only cover one direction.

3. Neighbour Entropy-Based Segmentation

All seven statistical models (Bigram, MI, t-model, boundary entropy, normalized MI, symmetric MI, and DCS models) only consider their segmentation benefits, making it hard to handle ambiguous words (sometimes combined with a lower benefit) and OOV words (sometimes occurring only once). This paper proposes a new model that considers neighbor entropies' segmentation benefits.

3.1. Entropy

In information theory, information entropy studies relate to the amount of information in a transmitted message. The definition of information entropy is expressed in terms of a discrete set of probabilities p_i [21]:

$$H = \sum_i h_i = \sum_i p_i \log_2 p_i \qquad (12)$$

For transmitted messages, these probabilities mean that a particular message was transmitted, and the entropy of the message system is a measure of the average amount of information in it. For medium and large-sized texts, the maximum probability of each character will be less than 0.1 in practice, and $h_i \propto p_i$ (see Figure 1). Thus, we use information entropy to calculate the benefit of the combined unit. The higher the probability is, the more significant is the entropy value.

3.2. Neighbour Entropy-Based Segmentation (NES)

When certain particles are combined, most available models only focus on the combination benefit rather than considering the loss caused by the combination (in other words, the benefit of separating them). From a statistical point of view, the combined unit should have a higher probability, while the unit may occur only once in terms of an OOV word. Therefore, there is a conflict in focusing on the unit's fitness value. In optimization research, the benefit function is established by itself and its "neighbors."

For each particle w_i, only two segmentation results are combined and separated with the continuous particle w_{i+1}, given by a label assignment of $label = \{Yes, No\}$. Here, $label = No$ if

there is a segment boundary between particles, and otherwise, *label = Yes* if there is a combination between particles. Based on this annotation, we define the benefit ratio of the text segmentation as follows:

$$Ratio_{benefit}(Yes \mid w_i) = \frac{Benefit(Yes \mid w_i)}{Benefit(Yes \mid w_i) + Benefit(No \mid w_i)} \quad (13)$$

where $Benefit(Yes \mid w_i)$ is the benefit of combining w_i and w_{i+1} together, $Benefit(No \mid w_i)$ is the benefit of separating w_i and w_{i+1} (or the loss of combination). If the entropy calculates the benefit functions, the benefit ratio equation can be rewritten as:

$$Ratio_{H}(Yes \mid w_i) = \frac{H(Yes \mid w_i)}{H(Yes \mid w_i) + H(No \mid w_i)} \quad (14)$$

The benefit value of $H(Yes \mid w_i)$ and the loss value of $H(No \mid w_i)$ are calculated by the following entropies:

$$H(Yes \mid w_i) = H(w_{i-2}, w_{i-1}) + H(w_{i+2}, w_{i+3}) \quad (15)$$

$$H(No \mid w_i) = H(w_{i-1}, w_i) + H(w_{i+1}, w_{i+2}) \quad (16)$$

$H(w_{i-2}, w_{i-1})$ and $H(w_{i+2}, w_{i+3})$ are calculated by Equation (12), referring to the entropy of two continuous particles, and are proportional to their probabilities. For example, $H(w_{i-2}, w_{i-1}) = -p(w_{i-2}, w_{i-1})\log_2 p(w_{i-2}, w_{i-1})$.

The segmentation process is illustrated in Figure 2. If w_i and w_{i+1} are combined, the probabilities are high to merging the former two characters (w_{i-2}, w_{i-1}) and following two characters (w_{i+2}, w_{i+3}). On the contrary, if w_i and w_{i+1} are separated, the probabilities are high for merging the former two characters (w_{i-1}, w_i) and following two characters (w_{i+1}, w_{i+2}). Four nearest neighbor entropies are included here to analyze the total benefits of segmentation.

$H(w_i, w_{i+1})$ is not included in $H(Yes \mid w_i)$, which means its benefit is not included, while the benefits of nearest neighbors are included, mainly due to the above-discussed conflict. Especially for the ambiguous words, the benefit value of $H(w_i, w_{i+1})$ is low. A bad combination could lead to losses on both sides of neighboring segmentations. Similarly, a lousy separation could result in losses on both sides of neighboring segmentations.

NES applies $Ratio_H(Yes \mid w_i)$ to segment the text, and is composed of four neighbor entropies, covering more information than other statistical models. Here, only bigrams are considered in NES calculations because trigram, 4-gram, and more are composed of bigrams.

3.3. Analysis of NES

NES is a measure of the benefit ratio between the combined condition and the separated con-

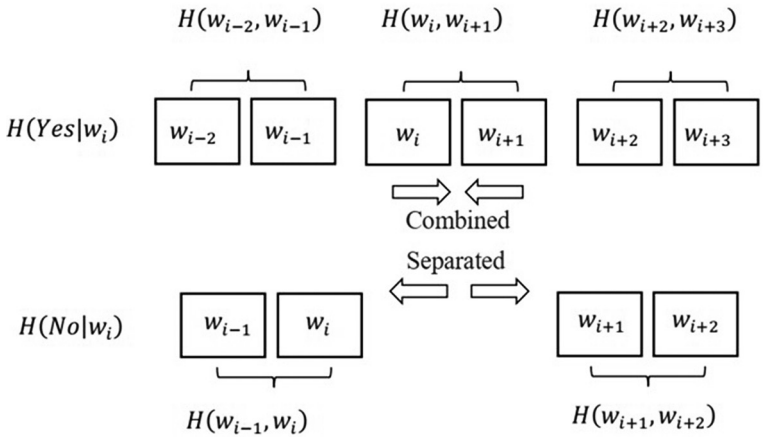

Figure 2. The neighbor entropy-based word segmentation process.

Table 2. An example of Chinese text segmentation of "不|符合".

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
	后	厨	临	时	用	电	不	符	合	安	全	要	求
NES	0.8	0.4	0.7	0.3	0.7	0.5	0.4	0.7	0.2	0.9	0.2	1.0	0.0

Note: 后厨临时用电不符合安全要求 (In the kitchen, temporary power usages do not meet safety requirements). '|' in the table means separators between Chinese characters.

Table 3. An example of Chinese text segmentation of "不符|合".

	c_1	c_2	c_3	c_4	c_5	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
	应	急	演	练	记	录	不	符	合	要	求
NES	0.7	0.2	0.8	0.2	0.8	0.4	0.53	0.4	0.5	0.8	0.0

Note: 应急演练记录不符合要求 (Emergency drill records do not meet the requirements).

dition, and four neighbor entropies ($H(w_{i-2}, w_{i-1})$, $H(w_{i+2}, w_{i+3})$, $H(w_{i-1}, w_i)$, $H(w_{i+1}, w_{i+2})$) are considered to contain more information than the traditional models that only cover $H(w_i, w_{i+1})$. The NES model shows three characteristic properties.

The first is the capacity of NES to measure any character. Equation (14–16) shows that NES is built from the nearest entropies and does not need extra statistics. The entropy value is zero if the neighbor variable is empty or unigram. For example, if there is no w_{i-1} or w_{i-2} for current w_i in one sentence, $H(w_{i-2}, w_{i-1}) = 0$ ($p(w_{i-2}, w_{i-1}) = 0$). No w_{i-1} means w_i is the beginning character of the sentence; no w_{i-2} means w_i is the second character of the sentence. The condition is the same for the last character of the sentence $H(w_{i+2}, w_{i+3})$.

The second is that the value of NES is between 0 and 1, and it is nondimensional, as shown in Equation (14). That is, $0 \le Ratio_H(Yes \mid w_i) \le 1$. The higher $H_N(Yes \mid w_i)$ is, the more significant is the probability of combining with a continuous one. Otherwise, 0 implies no benefit in combining particles. The smaller $H(No \mid w_i)$ is, the more significant is the probability to combine with a continuous one. The lowest value of NES appears at the word boundary.

The third is that the segmentation is dynamic for the exact text string. The segmentation may be different for the same text string based on the contextual relationship. It reflects the dynamic essence of word segmentation, and this paper uses the NES model as the objective function to segment the text into strings. At the same time, the segmentation result is not flexible for the exact text string in the test corpus for the other statistical models. For example, "不符合" (do not meet) can be segmented as "不|符合" and "不符|合," as shown in Table 2 and Table 3 (the character with the most significant NES value is combined first); in fact, both conditions ("符合" and "不符") are correct for CWS.

4. Maximum-Based Segmentation Algorithm

In order to verify the effectiveness of the NES model, a simple maximum-based segmentation algorithm is designed. One Chinese word can be composed of a unigram character, bigram characters, or more characters. *N*-gram characters ($N \ge 3$) can be composed of unigrams or bigrams, so identifying unigrams and bigrams is the fundamental process. The main idea of the max-algorithm is to segment unigrams and bigrams first by maximum fitness values in or-

der, and then combine them into *n*-grams. The steps of text segmentation are presented in the following bulleted list:

- input original text stream;
- pre-processing;
- initialize the stream with the binary vector, and segment the stream by punctuations;
- segment the stream by memorable characters (begin and end characters);
- segment the stream by maximum fitness values calculated by NES;
- combine into *n*-gram.

4.1. Segmentation Encodings

Since a Chinese sentence is usually a sequence of Chinese characters, we can take the CWS as a sequence tagging or labeling problem [38]. For each character in a sentence, one of the tags is assigned from a predefined tag set, indicating its segmentation result in a text stream. If a text stream contains n characters, the segmentation result of the text can be represented as a binary vector with n elements:

$$X = \{x_1, x_2, ..., x_i, ..., x_n\} \quad (17)$$

where x_i represents the segmentation value of the ith word in the text. It is encoded as either 0 or 1. If the ith word is combined with its next successive word, $x_i = 1$; otherwise, $x_i = 0$. Table 4 shows an example of the text segmentation result encoded by a binary vector.

4.2. Segmentation Steps

CWS consists of two parts. The first is pre-processing, and the second is word segmentation. The former extracts the valid characters from sentences with techniques including removing auxiliary words and segmenting sentences by punctuations. The second step segments the character series into words with meaningful units. The following steps elaborate on the segmentation method. Suppose $C = \{c_1, c_2, ..., c_i, ..., c_n\}$ is a sentence with n characters.

Step 1. *Init binary vector*. Init binary vector into −1 and 0. −1 means the following variable is a valid character, and 0 means the following variable is a punctuation; therefore, the binary value for the last character in a sentence is 0 (word boundary).

Step 2. *Identify the beginning and end characters*. Identify special characters that signify the beginning or end of the segmentation unit. For example, some characters such as "未 (no)" and "不 (no)" can be used as the beginning of words; some characters such as "等" (*etc.*), "及" (and), and "或" (or) can be used as the natural word boundaries of Chinese text. These characters are also called stop words, and there is a proposed stop word list [12]. In this step, the corresponding binary value should be set to 1 (for beginning words) or 0 (for stop words).

Step 3. *Segment by maximum*. This step aims to design an efficient algorithm for selecting a compact set of words. The most straightforward segmentation algorithm is setting a threshold; there should be a separator if the fitness value is less than the threshold. The smaller the fitness value, the higher the probability of separating them, and the bigger the fitness of two particles, the higher the probability of combining them. Based on this concept, we design a segmentation algorithm by ordering fitness values by maximum, combining the characters with the highest fitness value until there are no continuous unigrams. The task in this step is to segment an input sentence into a sequence of unigram and bigram characters. The advantage of this method is that users do not need to define the threshold relying on previous experience. Here, the corresponding binary value should be 1 if it meets the above requirements.

Step 4. *Combine into n-gram*. This step checks whether unigrams and bigrams can be combined into an n-gram ($n \geq 3$). If the fitness values of the remaining unhandled particles with the value "−1" are large enough, we will combine them. The threshold should be set by users depending on segmentation requirements.

Table 4. An example of Chinese text segmentation encoded by a binary vector.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
	后	厨	临	时	用	电	不	符	合	安	全	要	求
Binary vector	1	0	1	0	1	0	1	1	0	1	0	1	0

Note: 0 means the word boundary; 1 means the combination with the successive character.

Input (text stream)

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
后	厨	临	时	用	电	不	符	合	安	全	要	求
0.8025	0.3680	0.7041	0.2831	0.7131	0.4887	0.3872	0.7145	0.2335	0.9210	0.2311	0.9766	0.0000

Step 1. Inti binary vector

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	**0**

Step 2. Identify beginning and end characters

-1	-1	-1	-1	-1	**0**	-1	-1	-1	-1	-1	-1	0

Step 3. Segment by Maximum

-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	**1**	0
-1	-1	-1	-1	-1	0	-1	-1	-1	**1**	-1	1	0
1	-1	-1	-1	-1	0	-1	-1	-1	1	-1	1	0
1	-1	-1	-1	-1	0	-1	**1**	-1	1	-1	1	0
1	-1	-1	-1	**1**	0	-1	1	-1	1	-1	1	0
1	-1	**1**	-1	1	0	-1	1	-1	1	-1	1	0

Step 4. Combine into N-gram

1	**0**	1	**1**	1	0	**1**	1	**0**	1	**0**	1	0

Output (Segmentation)

后	厨	临	时	用	电	不	符	合	安	全	要	求
In kitchen		temporary		power usages		do not meet			safety		requirements	

Figure 3. An example of segmentation steps.

An example of segmentation steps of one sentence is shown in Figure 3. For example, in step 3, the most significant fitness value of the characters in this sentence is c_{12} "要." Therefore, we set x_{12} to 1 first. The second biggest fitness value is c_{10}, and therefore, we set x_{10} to 1. The fitness value of c_7 is also significant because the latter character's binary value is already 1, and therefore, we keep $x_7 = -1$ in step 3. Otherwise, there will be one trigram segmentation result in step 3. In step 4, if the count of "临时用电" is significant, we set $x_4 = 1$; otherwise, we set it to 0; moreover, as "不"(no) is the beginning character, we set $x_7 = 1$.

4.3 Measurements

Three segmentation measurements are used to evaluate the performance of an automatic segmentation method. These are precision (P), recall (R), and f-measure (F), which can be calculated as follows:

$$P = \frac{\text{correctNum}}{\text{autototalNum}} \tag{18}$$

$$R = \frac{\text{correctNum}}{\text{mutualCorrectNum}} \tag{19}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{20}$$

where "correctNum" is the number of words correctly identified by the automatic method, "autoTotalNum" is the total number of words identified by the automatic method, and "manualTotalNum" is the number of words identified

in the manual segmentation. A perfect segmentation algorithm will have these three functions at 100% [39].

Suppose there is a character stream for precision calculation, and the manual segmentation is "... $|c_{i-2}, c_{i-1}, | c_i, c_{i+1}, c_{i+2}, | c_{i+3}, c_{i+4} |$ ...". If the automated segmentation unit is "c_{i-1}, c_i" and "c_{i+2}, c_{i+3}" or "$c_i, c_{i+1}, c_{i+2}, c_{i+3}$," then the segmentation result is incorrect. If the segmentation unit is "c_i, c_{i+1}" or "$c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}$,", we ignore the segmentation result. The former causes a minimal segmentation unit, and the latter causes a huge segmentation unit. The segmentation result is correct only if the segmentation unit is "c_i, c_{i+1}, c_{i+2}." During precision calculation, the ignored results are not included in "autoTotalNum."

5. Case Study

Over the years, many companies and government agencies have been recording hazard texts in textual reports; however, this valuable knowledge is left unexploited because of the lack of suitable methodologies and high costs related to manual content analysis. Based on the model and the algorithm proposed in this paper, we extract meaningful units from semi-structured texts.

5.1. Database

In the national GDP, the catering industry share keeps increasing year by year. For example, in the first half of 2019 the catering industry revenue in China reached 2127.9 billion yuan, up 9.4% (http://www.gov.cn/xinwen/2019-07/15/content_5409256.htm), and the work safety in the catering industry warrants particular attention in urban operations, especially for large cities without industries. Therefore, the following experiments use the test corpus (see Appendix A) provided by a district government of the Beijing Administration of Work Safety, and the hazard texts are collected during the fourth quarter of 2018 for catering enterprises. There are 265 records and 4677 valid Chinese characters.

5.2. Experimental Results

First, a tool is used to segment the text. One of the most popular CWS tools is ICTCLAS (available at www.ictclas.nlpir.org), whose functions include word segmentation. However, like other popular tools, its segmentation results are not optimized for hazard texts. The majority of the segmented words are composed of unigram and bigram characters that are unnecessary for the hazard text. However, the hazard text is more likely to be segmented into more extended characters. For example, "后厨" (kitchen) is a meaningful unit and does not need to be segmented as "后|厨"; "不符合" (do not meet) is a compact unit that reflects the negative character of the hazard text and does not need to be segmented as "不|符合" (do not meet); and so on. Using ICTCLAS to analyze the test corpus, the tool achieves 99.3% precision, 80.1% recall, and 88.7% f-measure. The recall value is low, as only a limited number of auto-segmented units are composed of a bigram or more (29.7% unigram, 67.5% bigram, 2.8% trigram, and 0% others), while the manual segmentations contain several bigram and trigram characters (6.7% unigram, 79.2% bigram, 14.1% trigram, and 0% others). Therefore, some improvements are needed to realize better outputs to meet the end-user requirements.

There are some segmentation errors for the ICTCLAS tool. For example, "电源|线路" (power circuit) is wrongly segmented as "电源线|路", and "石油|气瓶" (gas cylinder) is wrongly segmented as "石油气|瓶." The tool depends on the dictionary approach, in which "电源线" (power cord) and "石油气" (oil gas) must be in the dictionary. In the future, we can establish a hazard text dictionary to improve the accuracy and the efficiency of hazard text segmentations.

Second, the text can be segmented using statistical models. Based on the model and the algorithm proposed in this paper, we developed an application in MATLAB R2016a, which extracts meaningful units from semi-structured hazard texts.

In order to compare the efficiency of the proposed model with other existing statistical models, we use the former three steps discussed in Section 4.2. Table 5 shows the performance of these models on the test corpus. The performance of boundary entropy is the

Table 5. Performance comparations for eight text segmentation models.

	Bigram	MI	t-model	Boundary Entropy	Normalised MI	Symmetric MI	DCS	NES
Precision	94.8%	95.2%	95.2%	94.7%	97.9%	96.8%	96.8%	99.3%
Recall	92.1%	92.6%	91.7%	90.6%	96.3%	94.5%	94.6%	98.7%
F-measure	93.5%	93.9%	93.4%	92.6%	97.1%	95.6%	95.7%	99.0%

worst because the model considers the conditional probability $p(w_{i+1} \mid w_i)$ only. The *t*-model is better because it covers two probabilities, $p(w_{i+1} \mid w_i)$ and $p(w_i \mid w_{i-1})$, considering the former and the latter relationships with the current character. Bigram probability and MI have the same performance as the t-model, and both models consider one co-occurrence probability parameter $p(w_i \mid w_{i+1})$. Normalized MI, symmetric MI, and DCS achieve high performances because these models cover two directions together. The models consider whether w_i would like to combine it with w_{i+1} and whether w_{i+1} would like to combine it with w_i. The NES model achieves the highest performance, while precision, recall, and f-measure are 99.3%, 98.7%, and 99.0%, respectively. NES covers four parameters during fitness calculation, and its segmentation results are accurate. The results show that the more information covered by a model, the more accurate the segmentation will be. The NES model containing four types of information achieves the highest *f*-measure of 99.0%, while the boundary entropy with one conditional probability achieves the lowest *f*-measure of 92.6%. Other models with two or three types of information achieve the median performances.

5.3. Typical Examples and Analysis

Table 6 selects three typical examples to compare the segmentation results of eight models. "董雨倩" (Dong Yuqian) is a person's name; "警示标示" (warning sign) is an ambiguous segmentation text string because there are several similar strings, such as "警示标志" (warning sign) and "指示标识" (indicator sign). Here, the count of "示标" is more than "警示"; "职业病防护" (occupational disease protection) is also hard to be segmented because the count of "业病" is the same as the count of "病防."

These examples show that the NES can correctly segment ambiguous words and OOV words. The named entity recognition (NER) works well for all these models; however, for ambiguous words such as "Example 2" and "Example 3," only the NES model performs well. The NES model considers combination benefits and separation benefits calculated by four neighbor entropies. Although NES has some advantages over n-gram probability, MI, and other statistical models, they are all based on statistics. For this reason, the statistical method is strongly dependent on an annotated corpus. Theoretically, a more significant amount of training data with higher quality annotation will bring about better performance of the text segmentation system.

5.4. Discussion

Mis-segmentation. There are some scenarios NES cannot handle, which causes its precision to be lower than 100%. First, $H(w_i, w_{i+1})$ is not included in the NES model, which causes some segmented units to be combined, though they have low frequencies. For example, "疏散|通道|存放|杂物" (The debris is stored in the evacuation escape route.) is wrongly segmented as "疏散|通|道存|放杂物," where "道存" only occurs once. We have tried several NES based improvements, but the current model is the best. Second, the maximum algorithm identifies unigram and bigram characters firstly. Therefore, it cannot correctly segment two continuous trigram characters, especially with lower occurrences. For example, "可调式|减压阀"

(adjustable pressure relief valve) is wrongly segmented as "可调|式减|压阀." Fortunately, these feature words with lower frequency have a negligible effect on information retrieval and text visualization.

Rules of manual segmentation. The manual segmentation of the database is an emotional problem. In Section 4.3, the standard segmentation is "... $|c_{i-2}, c_{i-1}, | c_i, c_{i+1}, c_{i+2}, | c_{i+3}, c_{i+4} |$ ..." that experts define. The rules of segmentation are listed as follows:

- separate conjunction characters as a unigram, such as "等" (*etc.*), "及" (and), "或" (or); moreover, most units are composed of bigrams;
- a negative character is combined with the following feature word, such as "不符合" (do not meet), "未健全" (need improve), "不正确" (incorrect)…, and these feature words reflect the characteristics of the hazard text;
- there are no 4-gram characters in the manual dataset to compare the effectiveness of the method with other segmentation models.

These rules are consistent with the former three segmentation steps in Section 4.2. The manual dataset is given in Appendix A. The correct segmentation is hard to reach an agreement on, but the failed segmentation is easy to be identified. Consequently, the segmentation results, including small and large units, are ignored during performance measurements. In order to compare the statistical models with fairly, 4-gram is not included in manual segmentations.

5.5. Information Retrieval and Text Visualization

After completing step 4 of the segmentation algorithm, the feature words will contain 4-gram

Table 6. Segmentation results of 8 models: Some examples and analyses.

		Example 1							Example 2				Example 3				
Input	Character	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
	Unigram count	20	45	36	25	1	1	1	7	57	45	57	9	45	6	46	28
	Bigram count	20	20	22	1	1	1	0	7	37	13	0	9	6	5	5	3
	Segmentation	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
		employee Dong Yuqian							warning sign				occupational disease protection				
Bigram	Fitness value	4.3E-03	4.3E-03	4.7E-03	2.1E-04	2.1E-04	2.1E-04	0.0E+00	1.5E-03	7.9E-03	2.8E-03	0.0E+00	1.9E-03	1.3E-03	1.1E-03	1.1E-03	6.4E-04
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
MI	Fitness value	2.9E-02	2.5E-02	3.2E-02	1.6E-03	2.6E-03	2.6E-03	0.0E+00	9.5E-03	4.8E-02	1.3E-02	0.0E+00	1.3E-02	8.6E-03	6.8E-03	4.5E-03	4.7E-03
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
t-model	Fitness value	4.47	-2.27	1.02	-4.19	0.96	0.00	-1.00	2.45	-0.89	-2.70	-3.61	2.72	-2.57	1.86	-1.93	-0.02
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
Boundary Entropy	Fitness value	0.00	-1.89	-1.77	-3.59	0.00	0.00	-∞	0.00	-0.86	-1.31	-∞	0.00	-1.89	-0.65	-1.95	-1.42
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
Normalised MI	Fitness value	0.85	0.46	0.80	0.62	1.00	1.00	0.00	0.68	0.75	0.20	0.00	0.74	0.70	0.56	0.10	0.70
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
Symmetric MI	Fitness value	0.58	0.42	0.68	0.08	1.00	1.00	0.00	0.21	0.68	0.18	0.00	0.32	0.22	0.18	0.08	0.19
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
Data Correlation	Fitness value	0.44	0.25	0.54	0.04	1.00	1.00	0.00	0.12	0.53	0.07	0.00	0.20	0.13	0.09	0.02	0.11
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护
NES	Fitness value	0.73	0.45	0.50	0.48	0.88	0.67	0.00	0.58	0.45	0.41	0.00	0.50	0.38	0.51	0.52	0.50
	Segmentation result	从	业	人	员	董	雨	倩	警	示	标	示	职	业	病	防	护

Table 7. Final segmentation results.

1-gram			2-gram			3-gram			4-gram		
Chinese	**English**	**Count**	**Chinese**	**English**	**Count**	**Chinese**	**English**	**Count**	**Chinese**	**English**	**Count**
无	no	38	后厨	kitchen	59	不符合	do not meet	38	安全生产	work safety	57
内	inside	25	培训	training	35	未健全	need improve	30	安全出口	emergency exit	47
和	and	15	缺少	lack of	33	未设置	do not set	16	堆放杂物	store debris	31
的	of	12	教育	education	30	防水型	waterproof	14	疏散通道	evacuation route	26
或	or	10	制度	system	25	照明灯	floodlight	12	从业人员	employees	20
及	and	10	记录	record	24	配电箱	power distribution box	9	临时用电	temporary power usage	19
被	passive	6	安全	safety	23	未如实	not truly	9	事故隐患	accident hazard	18
对	for	5	应急	emergency	22	防爆灯	explosionproof light	7	规范要求	requirements	18
等	Etc.	4	敷设	laying	22	新上岗	new hiring	7	指示标识	indication sign	15

characters. If the threshold is set to 8 (the minimum frequency of a 4-gram), the final result contains 7.5 % unigram, 57.9 % bigram, 16.6 % trigram, and 18.0 % 4-gram characters. The word segmentation results contain more bigram, trigram, and *n*-gram ($n \geq 4$) characters. Table 7 shows top 9 feature words for every *n*-gram (n =1, 2, 3, 4). These retrieved feature words can be served as the input data for text visualization. From Table 7, the result can be obtained as follows:

- there are many words on "安全生产" (work safety) and "事故隐患" (accident hazard), along with many negative words, which show it is a hazard report;
- the most unsafe places include "后厨" (kitchen), "安全出口" (emergency exit), and "疏散通道" (evacuation route); moreover, the most unsafe equipment includes "照明灯" (floodlight) and "防爆灯" (explosion-proof light);

- the riskiest persons are "新上岗" (new hiring), who lack "教育" (education) and "培训" (training); the most dangerous operations include "堆放杂物" (store debris) and "临时用电" (temporary power usage).

In summary, the segmented feature words retrieved from the original hazard text can help understand the safety risk, while the words with high frequencies reveal the common safety risks of catering enterprises.

6. Conclusion and Future Research Directions

6.1. Conclusion

The hazard text is a unique sub-category, containing 7.5 % unigram, 57.9 % bigram, 16.6 % trigram, and 18.0 % 4-gram or more characters. That means the general text segmentation methods need to be re-designed based on more characters in a segmentation unit.

The NES model proposed in this paper is valid for Chinese hazard text segmentation. It defines a benefit ratio of text segmentation, containing benefits and losses of combining the segmentation unit with more information (4 neighbor entropies) than other popular statistical models. It can also handle ambiguous and OOV words more efficiently. The model considers the segmentation benefits of neighbor entropies, adopting the concept of "neighbor" in optimization research. The experiments show that the performance of the boundary entropy is the worst; t-model, bigram probability, and MI show similar performances; normalised MI, symmetric MI, and DCS can achieve better results; while NES obtains the best results with a 99.3% precision, 98.7% recall, and 99.0% f-measure for Chinese hazard text segmentation. This paper also shows that the maximum-based segmentation algorithm can effectively handle the CWS problem in finding a highly compact subset from the original text stream at lower expenses. This algorithm first segments the text into unigrams and bigrams and then combines them into a meaningful n-gram ($n \geq 3$).

There are some limitations of the method. First, the self-benefit $H(w_i, w_{i+1})$ is not included in the NES model, causing some wrongly segmented units. We have tried several improvements based on NES, but the current model is the best. Second, the maximum algorithm identifies unigrams and bigrams firstly and cannot correctly segment two continuous trigram characters, especially with lower occurrences.

6.2. Industry Implications and Further Works

Hazard text is taken from a specific district government of the Beijing Administration of Work Safety. After text segmentation and statistical analysis, several common safety factors for catering enterprises are observed. For example, unsafe places include "kitchen," "emergency exit," and "evacuation route." Moreover, the most unsafe equipment includes "floodlight" and "explosion-proof light." The riskiest person is "new hiring," who lacks "education" and "training." The dangerous operations include "store debris" and "temporary power usage." In order to reduce risks, these common hazards should be added to the standardization of the hazard identification process for catering enterprises.

In the future, a critical task is to establish a dictionary library for safety science to make knowledge extraction more accurate and safety management more efficient.

Most importantly, the NES model can be verified for other languages such as English, as well as for other text segmentation problems. The simple statistical models are baseline methods for performance comparison; next, the neural network-based models can be designed for CWS on Chinese hazard texts, which have achieved satisfactory performance on various NLP tasks [19, 24, 25]. The hazard text's classification and clustering problem can be effectively handled based on extracted feature words.

Appendix

The dataset used in this study can be downloaded from https://github.com/qiao77/Hazard-Text-Catering/. Please cite this paper if the dataset is used.

Acknowledgment

This research was supported by the Beijing Key Laboratory of Megaregions Sustainable Development Modelling, Capital University of Economics and Business (CUEB) fund.

References

[5] P. Hughes *et al.*, "From Free-Text to Structured Safety Management: Introduction of a Semi-Automated Classification Method of Railway Hazard Reports to Elements on a Bow-Tie Diagram", *Safety Science*, vol. 110, pp. 11–19, 2018. http://dx.doi.org/10.1016/j.ssci.2018.03.011

[6] J. X. Liu *et al.*, "Neural Chinese Word Segmentation with Dictionary", *Neurocomputing*, vol. 338, pp. 46–54, 2019. http://dx.doi.org/10.1016/j.neucom.2019.01.085

[7] K. Vallmuur, "Machine Learning Approaches to Analysing Textual Injury Surveillance Data: A Systematic Review", *Accident Analysis and Prevention*, vol. 79, pp. 41–49, 2015. http://dx.doi.org/10.1016/j.aap.2015.03.018

[8] B. Brooks, "Shifting the Focus of Strategic Occupational Injury Prevention Mining Free-Text, Workers Compensation Claims Data", *Safety Science*, vol. 46, no. 1, pp. 1–21, 2008. http://dx.doi.org/10.1016/j.ssci.2006.09.006

[9] B. Ait Ben Ali *et al.*, "A Recent Survey of Arabic Named Entity Recognition on Social Media", *Revue d'Intelligence Artificielle*, vol. 34, no. 2, pp. 125–135, 2020. http://dx.doi.org/10.18280/ria.340202

[10] A. J. P. Tixier *et al.*, "Automated Content Analysis for Construction Safety: A Natural Language Processing System to Extract Precursors and Outcomes from Unstructured Injury Reports", *Automation in Construction*, vol. 62, pp. 45–56, 2016. http://dx.doi.org/10.1016/j.autcon.2015.11.001

[11] P. K. Nagaraj, *et al.*, "Kannada to English Machine Translation Using Deep Neural Network", *Ingénierie des Systèmes d'Information*, vol. 26, no. 1, pp. 123–127, 2021. http://dx.doi.org/10.18280/isi.260113

[12] C. Zhang *et al.*, "Text Sentiment Classification Based on Feature Fusion", *Revue d'Intelligence Artificielle*, vol. 34, no 4, pp. 515–520, 2020. http://dx.doi.org/10.18280/ria.340418

[13] Y. M. Goh and C. U. Ubeynarayana, "Construction Accident Narrative Classification: An Evaluation of Text Mining Techniques", *Accident Analysis and Prevention*, vol. 108, pp. 122–130, 2017. http://dx.doi.org/10.1016/j.aap.2017.08.026

[14] Z. Q. Cao, "Classification of Digital Teaching Resources Based on Data Mining", *Ingénierie des Systèmes d'Information*, vol. 25, no. 4, pp. 521–526, 2020. http://dx.doi.org/10.18280/isi.250416

[15] F. Zhang *et al.*, "Construction Site Accident Analysis Using Text Mining and Natural Language Processing Techniques", *Automation in Construction*, vol. 99, pp. 238–248, 2019. http://dx.doi.org/10.1016/j.autcon.2018.12.016

[16] S. Foo and H. Li, "Chinese Word Segmentation and Its Effect on Information Retrieval", *Information Processing & Management*, vol. 40, pp. 161–190, 2004. http://dx.doi.org/10.1016/S0306-4573(02)00079-1

[17] N. W. Chi *et al.*, "Using Ontology-Based Text Classification to Assist Job Hazard Analysis", *Advanced Engineering Informatics*, vol. 28, no. 4, pp. 381–394, 2014. http://dx.doi.org/10.1016/j.aei.2014.05.001

[18] F. Hogenboom *et al.*, "A Survey of Event Extraction Methods from Text for Decision Support Systems", *Decision Support Systems*, vol. 85, pp. 12–22, 2016. http://dx.doi.org/10.1016/j.dss.2016.02.006

[19] Z. Wu and G. Tseng, "Chinese Text Segmentation for Text Retrieval: Achievements and Problems", *Journal of the American Society for Information Science Banner*, vol. 44, no. 9, pp. 532–542, 1993. http://dx.doi.org/10.1002/(SICI)1097-4571(199310)44:9<532::AID-ASI3>3.0.CO;2-M

[20] M. Zhang *et al.*, "A Chinese Word Segmentation Based on Language Situation in Processing Ambiguous Words", *Information Sciences*, vol. 162, pp. 275–285, 2003. http://dx.doi.org/10.1016/j.ins.2003.09.010

[21] Q. Qiu *et al.*, "DGeoSegmenter: A Dictionary-Based Chinese Word Segmenter for the Geoscience Domain", *Computers & Geosciences*, vol. 121, pp. 1–11, 2018. http://dx.doi.org/10.1016/j.cageo.2018.08.006

[22] X. R. Wang, "Topical n-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval", in *Proceedings of the seventh IEEE international conference on data mining, ICDM*, 2007, pp. 697–702.

[23] H. C. Peng, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005. http://dx.doi.org/10.1109/TPAMI.2005.159

[24] Z. Q. Sun and Z. H. Deng, "Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modelling", in *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, Brussels. Association for Computational Linguistics, 2018, pp. 4915–4920.

[25] J. Wei *et al.*, "A Hybrid Linear Text Segmentation Algorithm Using Hierarchical Agglomerative Clustering and Discrete Particle Swarm Optimization", *Integrated Computer-Aided Engineering*, vol. 21, pp. 35–46, 2014. http://dx.doi.org/10.3233/ICA-130446

[26] J. Qiao and Y. Li, "Resource Leveling Using Normalized Entropy and Relative Entropy", *Automation in Construction*, vol. 87, pp. 263–272, 2017. http://dx.doi.org/10.1016/j.autcon.2017.12.022

[27] J. B. Xie *et al.*, "Chinese Alt Text Writing Based on Deep Learning", *Traitement du Signal*, vol. 36, no. 2, pp. 161–170, 2019. http://dx.doi.org/10.18280/ts.360206

[28] X. Chen *et al.*, "Long Short-Term Memory Neural Networks for Chinese Word Segmentation", in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Association for Computational Linguistics*, 2015, pp. 1385–1394.

[29] L. J. Zhao *et al.*, "Neural Networks Incorporating Unlabeled and Partially-labeled Data for Cross-domain Chinese Word Segmentation", in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018. http://dx.doi.org/10.24963/ijcai.2018/640

[30] M. Liu and P.C. Liao, "Integration of Hazard Rectification Efficiency in Safety Assessment for Proactive Management", *Accident Analysis and Prevention*, vol. 129, pp. 299–308, 2019. http://dx.doi.org/10.1016/j.aap.2019.05.020

[31] J. J. Lin and A.S. Morse "Coordination of Groups of Mobile Autonomous Agents Using Nearest Neighbor Rules", *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003. http://dx.doi.org/10.1109/tac.2003.812781

[32] Y. Z. Liang *et al.*, "Out-Domain Chinese New Word Detection with Statistics-Based Character Embedding", *Natural Language Engineering*, vol, 25, no. 2, pp. 239–255, 2019. http://dx.doi.org/10.1017/S1351324918000463

[33] Y. Doval and C. Gomez-Rodriguez, "Comparing Neural- and N-Gram-Based Language Models for Word Segmentation", *Journal of the Association for Information Science and Technology*, vol. 70, no. 2, pp.187–197, 2019. http://dx.doi.org/10.1002/asi.24082

[34] L. M. Abualigah and A.T. Khader, "Unsupervised Text Feature Selection Technique Based on Hybrid Particle Swarm Optimization Algorithm with Genetic Operators for the Text Clustering", *Journal of Supercomputing*, vol, 73, no. 11, pp. 4773–4795, 2017. http://dx.doi.org/10.1007/s11227-017-2046-2

[35] M. H. Bokaei *et al.*, "Extractive Summarization of Multi-Party Meetings Through Discourse Segmentation", *Natural Language Engineering*, vol. 22, no. 1, pp. 41–72, 2016. http://dx.doi.org/10.1017/S1351324914000199

[36] H. X. Fei *et al.*, "Chinese Word Segmentation Research Based on Statistic the Frequency of the Word", *Computer Engineering and Applications*, 2005.

[37] Z. Zhang *et al.*, "Toward Unsupervised Protocol Feature Word Extraction", *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 10, pp. 1894–1906, 2014. http://dx.doi.org/10.1109/JSAC.2014.2358857

[38] P. A. Estevez *et al.*, "Normalized Mutual Information Feature Selection", *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp.189–201, 2009. https://ieeexplore.ieee.org/abstract/document/4749258

[39] G. Karakaya *et al.*, "Identifying (Quasi) Equally Informative Subsets in Feature Selection Problems for Classification: A Max-Relevance Min-Redundancy Approach", *IEEE Transactions on Cybernetics*, vol. 46, no. 6, pp. 1424–1437, 2016. https://ieeexplore.ieee.org/abstract/document/7150365

[40] S. Song *et al.*, "Probabilistic Correlation-Based Similarity Measure on Text Records", *Information Sciences*, vol, 289, pp. 8–24, 2014. http://dx.doi.org/10.1016/j.ins.2014.08.007

[41] G. Qu *et al.*, "A New Dependency and Correlation Analysis for Features", *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1199–1207, 2005. http://dx.doi.org/10.1109/TKDE.2005.136

[42] N. W. Xue and L.B. Shen, "Chinese Word Segmentation as LMR Tagging", in *Proceedings of the 2nd SIGHAN workshop on Chinese language processing*, 2003, pp. 176–179.

[43] D. Zeng *et al.*, "Domain-Specific Chinese Word Segmentation Using Suffix Tree and Mutual Information", *Information Systems Frontiers*, vol. 13, no. 1, pp. 115–125, 2011. http://dx.doi.org/10.1007/s10796-010-9278-5

Received: April 2022
Revised: June 2022
Accepted: June 2022

Contact addresses:
Jianfeng Qiao
Capital University of Economics and Business
Beijing Key Laboratory of
Megaregions Sustainable Development Modeling
Beijing
China
e-mail: qiaojianfeng@cueb.edu.cn

Xingzhi Yan
University of Birmingham
UK
e-mail: yanxingzhi@cnpc.com.cn

Shuran Lv
Capital University of Economics and Business
Beijing
China
e-mail: lsr22088@cueb.edu.cn

Jianfeng Qiao is an associate professor at the Capital University of Economics and Business (CUEB), Beijing, China. He works in the Institute of Management and Engineering of CUEB. His research is in text mining, project management, system engineering, risk management, safety management, information entropy, modeling, and simulation. He published several research papers in system simulation, validation of simulation model, resource management, optimal control, risk identification, and risk evaluation. He published in journals that include Automation in Construction, Journal of System Simulation, and Computer Simulation. He completed several research projects, such as Safety Risk Management in Large Scale Engineering, Data Mining in Accident Narratives, and Study and Judgment of Safety Situation for Government.

Xingzhi Yan is a graduate student in the Computer Science School, at the University of Birmingham, Birmingham, UK. His research is in computer science, natural language processing, network security, safety management, project management, and risk management. He participated in several research projects, such as Nature Language Processing of Hazard Text, Fire Vulnerability Analysis and Visualization, Risk Identification and Analysis.

Shuran Lv is a professor at Capital University of Economics and Business (CUEB), Beijing, China. He works in the Institute of Management and Engineering of CUEB. His research is in risk management, monitoring and early warning, safety management, accident prevention, explosion protection, and mine safety. He participated in more than 100 scientific research projects from the National Natural Science Foundation of China, the Emergency Department of China, and the Natural Science Foundation of Beijing. He published several academic books, such as Prevention and Control of Safety Accidents and Case Analysis, Fire and Escape Simulation, Building Fire Simulation Engineering Software, and Fire and Explosion Prevention Technology.