ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

The Design and Implementation of an End-side Dictionary-based Chinese-English Word Segmentation Algorithm

Gao Qun[1,a,*]

[1]*School of Intelligent Transportation Modern Industry, Anhui Sanlian University, Hefei, 230601, China*
[a]*95219180@qq.com*
[*]*Corresponding author*

Abstract: *By analyzing the Chinese-English word segmentation requirements on the device side, this paper elaborates on the design and implementation process of a dictionary-based Chinese-English word segmentation algorithm on the device side. Through in-depth analysis of the word segmentation requirements on the device side and combining advanced algorithm strategies, an efficient and accurate word segmentation algorithm is designed. Verified by experiments, this algorithm performs well on different devices. On the manually annotated Chinese and English test sets, the accuracy of whole-sentence word segmentation reaches more than 91.3% and 82.1% respectively, providing an implementation idea for the implementation on the device side.*

Keywords: *Device Side; Chinese-English Word Segmentation Algorithm; Good Performance*

1. Introduction

1.1. Research Background and Significance

Various intelligent devices have become important tools for efficient work and improving learning quality. Especially learning auxiliary device types of equipment provide students with convenient learning methods and greatly enrich learning resources and environments. As a basic technology for text processing, Chinese-English word segmentation algorithms play a crucial role in improving the text recognition and processing capabilities of intelligent devices.

Through the research on Chinese-English word segmentation algorithms and optimization according to the characteristics of devices, it can not only promote technological development but also directly promote the performance improvement of algorithms on device-side equipment. In view of this, the Chinese-English word segmentation algorithm is studied, and an algorithm design idea and implementation suitable for device-side application scenarios is proposed. This algorithm will consider the special needs of terminal devices such as processing speed, user interaction experience, and accuracy, and strive to provide users with more accurate and convenient learning experiences in the constantly advancing technological trend.

1.2. Review of Related Work

In the research field of Chinese-English word segmentation algorithms, scholars have proposed various methods to improve the accuracy and efficiency of word segmentation. For Chinese, dictionary-based word segmentation methods, statistical-based models, and in recent years, popular deep learning methods[1-3]. In English word segmentation, in addition to the traditional word segmentation method based on spaces and punctuation marks, there are also word segmentation algorithms based on natural language processing, such as methods based on regular expressions and methods based on hidden Markov models (HMM)[4].

In the application of learning auxiliary equipment, word segmentation algorithms are mainly used to improve the text scanning and translation quality of equipment. For example, in scanning dictionary pens, word segmentation algorithms can correct the OCR text recognition results of user-scanned text, perform effective word segmentation so that users can click one by one and match with built-in resources, and can also assist subsequent speech synthesis and text translation. In intelligent education software, word

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

segmentation algorithms can help the software better understand and process user-input queries.

Although existing word segmentation algorithms have achieved certain application results in learning auxiliary equipment, there are still some challenges and deficiencies. For example, for sentences in complex contexts or rare words, the word segmentation effect of existing algorithms is still not ideal. In addition, one of the important usage scenarios of learning auxiliary equipment may be a network-free environment. In this scenario, the word segmentation algorithm can only use a local solution, and the requirement for real-time performance is high, while existing algorithms still have room for improvement in processing speed. Therefore, through in-depth analysis, the requirements of Chinese-English word segmentation algorithms in the learning auxiliary equipment scenario are studied, and a practical local Chinese-English word segmentation algorithm implementation plan is proposed to achieve better application effects in the learning assistance field.

1.3. Research Objectives and Contributions

The main goal of this research is to design and implement an efficient and feasible learning auxiliary device mainly implemented in pure C language for the special application scenario of Chinese-English word segmentation algorithms for learning auxiliary devices on the device side, aiming for better cross-platform and portability. Expected results include improving the accuracy of whole-sentence word segmentation, the time-consuming of each word segmentation, and assisting in improving the user experience (assisting in subsequent text translation and speech synthesis of learning auxiliary devices). By optimizing the algorithm, it is hoped to improve the performance of learning auxiliary devices in text analysis without increasing the burden on the system. In addition, this research will also explore the actual performance of the algorithm on two hardware platforms and verify its effectiveness and feasibility through algorithm performance and effect tests.

The contributions of the proposed scheme in this research are mainly reflected in the following three aspects: First, a new Chinese-English mixed text word segmentation algorithm is proposed, which considers the actual application situation in learning auxiliary devices, such as the requirements for real-time performance and accuracy. Secondly, the advantages of the proposed algorithm in accuracy and efficiency are verified through comparative experiments, providing more reliable text processing technical support for learning auxiliary devices. Finally, the algorithm implementation proposed in this research provides valuable references and inspirations for the implementation of Chinese-English word segmentation algorithms on learning auxiliary devices.

2. Introduction of Word Segmentation Algorithm

2.1. Overview of Chinese Word Segmentation Technology

Chinese word segmentation is a basic task in the field of natural language processing. Its goal is to cut a continuous sequence of Chinese characters into meaningful words. Since there are no obvious boundary marks between words in Chinese text, this task is extremely challenging for computers. Traditional Chinese word segmentation methods are mainly based on dictionary matching, including forward maximum matching, backward maximum matching and bidirectional matching. These methods are simple and effective, but they do not handle unregistered words and ambiguous segmentation well. With the development of technology, statistical learning-based models, conditional random fields and popular deep learning methods in recent years have been widely used in Chinese word segmentation tasks, significantly improving the accuracy of word segmentation.

2.2. Overview of English Word Segmentation Technology

The English word segmentation task is relatively simpler compared to Chinese word segmentation because English words are generally separated by spaces. However, English word segmentation is not without challenges, especially when dealing with derived words, abbreviations, numbers and special symbols. English word segmentation needs to consider factors such as stemming, part-of-speech tagging and syntactic structure. Commonly used technologies include regular expression matching, dynamic programming algorithms and machine learning-based methods. In recent years, the introduction of deep learning technology has further improved the performance of English word segmentation, especially in application scenarios dealing with irregular words and context-dependent situations.

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

2.3. Special Requirements of Intelligent Devices for Word Segmentation Algorithms

Intelligent devices, especially learning auxiliary devices, put forward special requirements for word segmentation algorithms. The primary task of Chinese-English word segmentation algorithms on learning auxiliary devices is to complete dictionary-based word segmentation. The dictionary is the built-in dictionary resource of learning auxiliary devices. Only when accurate word segmentation can be achieved can tasks such as scanning and looking up words be completed more quickly. First of all, accuracy is the most important consideration factor because incorrect word segmentation results will directly affect the learning effect. Secondly, real-time performance is also crucial. The device needs to return results within a limited time, which requires the word segmentation algorithm to have efficient processing capabilities. Therefore, for the application of learning auxiliary devices on the device side, dictionary-based word segmentation algorithms are mainly selected for design and implementation to meet the specific needs of learning auxiliary devices.

3. Algorithm Design

3.1. Algorithm Framework

According to the application scenario of learning auxiliary devices, the algorithm adopts the design idea of dictionary-based word segmentation algorithm. This algorithm matches the string to be matched with the words in an established "sufficiently large" dictionary resource according to a certain strategy. If a certain entry is found, it means that the matching is successful and the word is recognized. The specific algorithm strategy adopts the configuration method of forward maximum value of strings. The algorithm framework is shown in Figure 1 below.

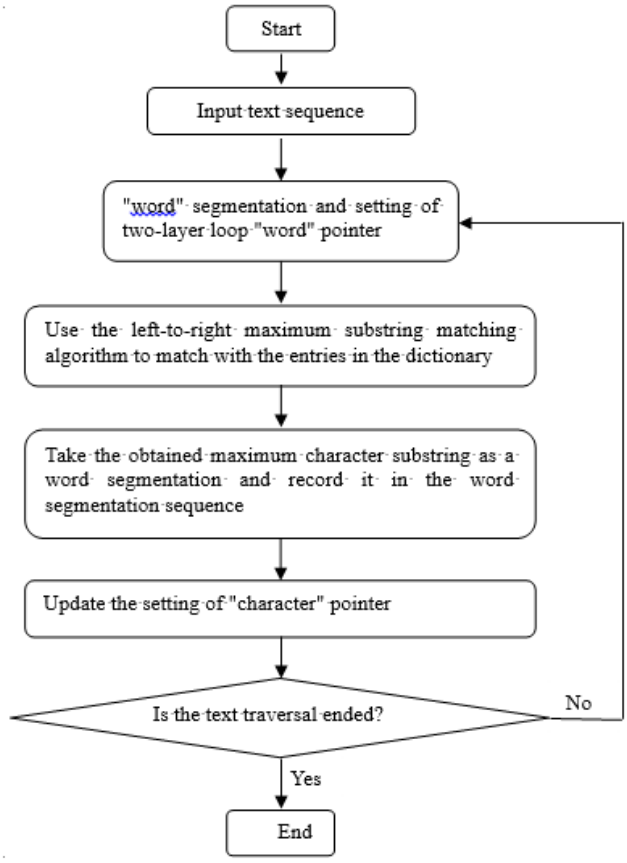

Figure 1: Flow chart of word segmentation

3.2. Detailed Description of Key Technologies

In the design of Chinese-English word segmentation algorithms, the key technologies mainly include the following aspects:

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

1) "word" segmentation and pointer setting: In this step, the algorithm needs to perform fine-grained segmentation of the text sequence, identify individual Chinese characters, punctuation marks, continuous letter strings and number strings. Then, set up a two-layer "word" pointer to track the current processing position and state.

2) Substring matching and dictionary lookup: This is the core step of the word segmentation algorithm. Starting from the position pointed to by the two-layer "word" pointer, according to the principle of forward maximum matching from left to right, intercept substrings in turn and match them with the entries in the dictionary. The dictionary lookup uses the hash algorithm to improve the lookup efficiency.

3) Substring recording and pointer update: When a matching entry is found, record it in the word segmentation sequence and update the two-layer "word" pointer to prepare for the next substring matching.

4) End condition check: After each loop ends, check whether the text string has been completely matched. If there is still an unfinished part, return to step 2 and continue execution; otherwise, the entire word segmentation process ends.

5) Dictionary construction: The dictionary is an indispensable part of the word segmentation algorithm, and it contains all possible words. When constructing the dictionary, it is necessary to consider the diversity and coverage of vocabulary, and it is also necessary to update the dictionary regularly to adapt to language changes.

6) String matching algorithm: The string matching algorithm is another important link in the word segmentation process. The hash algorithm is used for string matching in this study. This algorithm quickly locates and compares strings by converting strings into unique hash values, thereby improving matching efficiency and accuracy.

Through the application of the above key technologies, the Chinese-English word segmentation algorithm proposed in this study can effectively process text sequences in learning auxiliary devices and meet the requirements of high precision and high efficiency.

3.3. Algorithm Optimization Strategies

For the above algorithm framework and key technology descriptions, the following optimization strategies can be adopted to further improve the performance of the word segmentation algorithm:

1) Dictionary structure optimization: Optimize the data structure of the dictionary. For example, using a tree structure to organize the dictionary can greatly improve the string search efficiency.

2) Algorithm parallelization: Take advantage of the multi-core processors of modern computers to parallelize the computationally intensive parts of the algorithm, such as parallel string matching, to reduce the overall processing time.

3) Adaptive word segmentation strategy: Dynamically adjust the word segmentation strategy according to different application scenarios and user habits. For example, for more professional materials, use more refined word segmentation rules; for general materials, use standard word segmentation rules.

4) Feedback mechanism: Introduce a user feedback mechanism to continuously adjust and optimize the algorithm according to the actual usage of users. The dictionary resource packaging supports forced setting of word segmentation to provide a fallback for the algorithm and quickly respond to user needs. This includes collecting user feedback on word segmentation results, analyzing error cases, and continuously improving the word segmentation dictionary and word segmentation rules[5].

4. Algorithm Implementation

4.1. Dataset Preparation

According to the above word segmentation scenario requirements, the dataset is mainly divided into two parts:

4.1.1. Dictionary

According to the usage scenario of learning auxiliary devices, mainly scanning Chinese or English

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

textbooks to realize the function of looking up words.

The dictionary mainly collects some words and sentences from ancient poems in textbooks at the K12 education stage. The specific dictionary data is shown in Table 1 below.

Table 1: Dictionary data information

Dictionary dataset	number of sentences
Chinese characters (with word frequency)	22500
Chinese poetry, words, songs, classical Chinese	8500
Chinese pinyin and pinyin splitting	3150
English phrases	100500
Punctuation marks	380

4.1.2. Test Set

Referring to the framework and annotation methods of some other word segmentation test set corpora, some sentences from Chinese and English textbooks for primary, junior high and high schools are selected as the effect verification and evaluation set. The data annotation is carried out according to people's subjective perception, as shown in Table 2 below.

Table 2: Classification and quantity of test sets

Dataset	Quantity
Chinese primary, junior high and high school textbook sentences	13600
English primary, junior high and high school textbook sentences	24000

4.2. Implementation Environment and Tools

To ensure the wide applicability and portability of the algorithm, this study uses platform-independent C language to implement the Chinese-English word segmentation algorithm. C language has the characteristics of high efficiency, stability and cross-platform, making it an ideal choice for processing strings and implementing complex logic. In terms of hardware environment, the algorithm can run in any environment that supports C language compilation. The development of learning auxiliary devices generally uses C/C++ programming languages for development, which can meet the implementation requirements of learning auxiliary devices for the algorithm.

4.3. Implementation Details

4.3.1. Dictionary resources

The design quality of the data structure of dictionary resources will directly affect the algorithm performance. In order to improve the initialization efficiency of the word segmentation engine, the generation of dictionary resources should be implemented by offline generation and packaging into binary resource files[3], and then loaded into memory when the word segmentation engine is initialized.

To facilitate matching with different dictionaries in different application scenarios, it is necessary to support multiple different types of dictionaries.

The data structure design of each type of dictionary is as follows:

```
typedef struct {
    char    name[64];       //Dictionary name
    uint_t  word_max_bytes; //Maximum number of bytes for words in this dictionary
    char    longestword[128];//Longest word
    uint_t  length;         //Number of allocated tables
    uint_t  size;           //Number of actually used in length
    uint_t  factor;         //Control parameter for allocating length
    uint_t  threshold;      //Control parameter for allocating length
}ws_dict;
```

The above table data structure is as follows (12 bytes):

```
typedef struct {
    char *  word;       //Pointer to text string
    void *  val;        //the node pointer to struct lex_entry
    struct ws_dict * next;//Point to the next word with the same hash value
}ws_tab;
```

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

The above val data structure is as follows (16 bytes):

```
typedef struct {
    char    len;        //Length of entry
    char    type;       // Word type
    uint_t  offset;     //Number of offset bytes in the text
    char *  word;       //Pointer to the word text
    uint_t  freq;       //Word frequency, only available in Chinese character table.
}lex_desc;
```

4.3.2. Hash Algorithm for String Matching

Considering the subsequent word segmentation search speed, string hashing algorithm is used for storage for each entry. In this way, the relationship between the efficiency of word segmentation and the number of entries is not so great. It is allowed that different strings have the same hash value. In actual testing, this situation exists but is not frequent. (If memory overhead is not considered, the range of the hash-mapped array can be made larger, which can reduce hash value conflicts).

5. Experiment and Result Analysis

5.1. Results Related to Algorithm Efficiency

The algorithm supports input text in UTF8 and GBK formats. The longest single word segmentation is 1000 bytes (about 330 Chinese characters). The longest text in the test set is 988 bytes. In order to compare the performance of the word segmentation algorithm, two platforms were selected for testing. The test results are shown in Table 3 below.

Table 3: Test results of algorithm efficiency

Target platform	Processor	Memory required by algorithm	Maximum time-consuming for word segmentation (unit: millisecond)
PC	12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz	35MB	2
V853	Arm Cortex-A7 CPU core @ 1 GHz	35MB	43

5.2. Results Related to Algorithm Accuracy

The test set uses manually annotated word segmentation results. The accuracy of the algorithm is calculated in the following way. If the word segmentation result of the whole sentence is consistent with the manual annotation, the sentence is recorded as correct word segmentation. As long as there is a word segmentation error at one place, it is recorded as a word segmentation error. Word segmentation accuracy = number of correct word segmentation sentences / total number of test sentences. The specific test results are shown in Table 4 below.

Table 4: Word segmentation effects of each test set

Dataset	Quantity	Accuracy
Chinese primary, junior high and high school textbook sentences	13600	91.3%
English primary, junior high and high school textbook sentences	24000	82.1%

5.3. Result Analysis

This experiment has conducted a detailed evaluation of the efficiency and accuracy of Chinese-English word segmentation algorithms. Through algorithm performance testing on different platforms and accuracy verification on multiple data sets, the following analysis results can be obtained:

5.3.1. Analysis Related to Algorithm Efficiency

The tests of the algorithm on two different hardware platforms show that on a general PC, the time-consuming for word segmentation of an input text of 330 characters is only 2 milliseconds, which benefits from its strong computing power and high processor frequency. For the V853 chip, although it is equipped with a CortexA7@1 GHz processor, its word segmentation time-consuming significantly increases to 43 milliseconds. This shows that the performance of the algorithm on chips with weaker

ISSN 2706-655X Vol.6, Issue 6: 1-7, DOI: 10.25236/IJFET.2024.060601

processing capabilities needs to be further optimized. However, the overall performance is relatively good; in addition, the algorithm's memory usage is independent of the platform. In the case of supporting 500,000 dictionaries, only 35MB of memory is required, showing good memory management.

5.3.2. Analysis Related to Algorithm Accuracy

In the accuracy test, manually annotated word segmentation results were used as the benchmark to ensure the accuracy and authority of the evaluation. The test results show that on the Chinese primary, junior high and high school textbook sentence data set, the word segmentation accuracy reaches 91.3%, while on the English primary, junior high and high school textbook sentence data set, the accuracy is slightly lower, at 82.1%. The overall algorithm effect is usable.

The performance on the English test set is not as good as that on the Chinese test set. After analysis, there are mainly two reasons: First, there are some problems with the annotation of the test set. Because the annotators are not as familiar with English word segmentation as they are with Chinese, there is a large individual subjectivity difference in the annotation results. For example, for the sentence “This is the way I go to bed.”, it can be segmented into “This is/the way/I/go to bed/./” and “This/ is/the/way/I/go to bed/./”, but when calculating the accuracy rate, it is compared with the annotation, which may lead to a low accuracy rate statistics; second, there are some special cases in English word segmentation, such as hyphens, abbreviations such as I’m, etc., which need to be further optimized.

Based on the above analysis, the Chinese-English word segmentation algorithm in this study shows efficient and relatively accurate word segmentation capabilities. However, there is still room for improvement in the optimization for low-performance processors and the accuracy of English text word segmentation. Future work can focus on improving the algorithm to adapt to low-performance hardware environments and refining and optimizing the processing rules for English texts.

6. Conclusion and Outlook

An in-depth research and practice on the Chinese-English word segmentation algorithm in learning auxiliary devices has been conducted, and an efficient and accurate word segmentation method has been proposed and detailed design and implementation have been carried out for this method. Tests on two platforms show that this word segmentation algorithm has a fast processing speed while ensuring high accuracy, and can meet the real-time requirements of learning auxiliary devices.

Through experiments, it can also be seen that there is still room for improvement in the current work. Especially in the processing of English texts, how to further improve the accuracy of word segmentation and how to deal with special language phenomena such as hyphens and abbreviations will be the focus of future work. In addition, optimization for low-end hardware devices is also one of the future research directions, hoping that this algorithm can be effectively applied on a wider range of learning auxiliary devices.

Acknowledgements

This work was supported by the University Scientific Research Fund Project No. PTZD2024013.

References

[1] Chao Shen, Jinxia Dai, Mengge Mao, et al. Research on Chinese Word Segmentation Algorithm based on Dictionary and Statistical Method and Its Application in the field of Power Grid Control[C]//3rd International Symposium on Information Science and Engineering Technology(siset2022), 2022: 19-23.

[2] Huang Linjieqiong, Li Xingshan. The effects of lexical- and sentence-level contextual cues on Chinese word segmentation. [J]. Psychonomic Bulletin & Review, 2023: null.

[3] Liu Yang, Yu Tian, Ding Yi. A New Chinese Word Segmentation Based on Maximum Probability Path [J]. Computer & Digital Engineering, 2022, 50(03):591-596.

[4] Zou Zhimin, Guo Heqing, Gao Ying. English String Segmentation Method[J].Application Research of Computers, 2007, (07):52-54.

[5] Zhang Jun, Lai Zhipeng, Li Xue. Cross-domain Chinese Word Segmentation Based on New Word Discovery[J].Journal of Electronics & Information Technology, 2022, 44(09):3241-3248.