A Novel Spell Checking Algorithm for Non-Segmented Languages

Tim Hunt
Waikato Institute of Technology
Tristram Street
Hamilton
+648348800 ext 8726
tim.hunt@wintec.ac.nz

Blaine Rakena
Waikato Institute of Technology
Tristram Street
Hamilton
+648348800 ext 8726
Blaine.Rakena@wintec.ac.nz

Kevin Wang
Waikato Institute of Technology
Tristram Street
Hamilton
+648348800 ext 8726
Kevin.wang@wintec.ac.nz

ABSTRACT
This work describes for the first time a novel approach to implementing a successful spell checker for non-segmented languages such as Chinese. The design combines a novel technique that we call First Character Identifier Approach (FCIA). FCIA elegantly side steps the widely reported problem of how to segment text into words in a timely fashion. It does this by using a list of incorrectly spelt words to search for word matches in the text using the first character of the incorrect word. The design has been successfully implemented into a children's email application that can be downloaded from http://www.mifrenz.com.

Categories and Subject Descriptors
F.2.2 [**Nonnumerical Algorithms and Problems]**: Pattern matching.

General Terms
Algorithms, Performance, Design.

Keywords
Non-segmented languages, spell checker, Mifrenz.

1. INTRODUCTION
An accurate spell checker is a part of most modern word processing applications. Although this has long been the case for languages that delimit the words in a sentence with a space, there is still on-going research into how best to segment sentences of languages that do not use spaces between words such as Chinese, Thai and Japanese.

This work describes for the first time an original design, development and implementation of a spell checking algorithm for the Chinese language called First Character Identifier Approach (FCIA). A critical point of difference of FCIA from other spell checkers is that the issue of segmentation is circumvented by designing a simple approach of using the sentence characters as keys to look up a dictionary of words. This approach is combined with the use of a dictionary of incorrect rather than correct words, something that other studies have used, but taken together results in a novel approach that is fast simple and robust. The algorithm was implemented using the Java programming language and integrated into the Mifrenz email application for children [2].

2. THE PROBLEM
One of reasons behind the lag of the development of a Chinese language spell checker is complexity caused by two facts: the first is that Chinese text has no natural delimiters such as spaces between words and it is possible for a sentence to be segmented in several legitimate ways [4], the second is that there is the problem with unknown words [5] which are not predefined in a dictionary but can be created by combining characters or words.

2.1 Segmentation
The widely used approaches in creating a Chinese spell checker usually include two steps: 1) work out an algorithm to segment the text and then 2) detect any errors. Grammatical rules are also sometimes used to detect errors [4]. The first obstacle people encountered in finding a solution for a Chinese spell checker was segmentation, that is identifying the correct combinations of the Chinese characters in a sentence or, in other words, finding a way to correctly group the characters in a sentence. The complexity of segmentation leads to a very large number of segmentation possibilities, which in turn gives the problem of how to examine all those possible segmentations efficiently, a task yet to be solved. The number of possible segmentations (N) for a sentence of length (L), with a maximum word length (m) is given as:

$$N_L = \sum_{i=1}^{i=m} N_{L-i} \quad \text{Equation 1.}$$

where $N_{L-i} = 0$ if $L - i < 0$; and $N_{L-i} = 1$ if $L = i$.

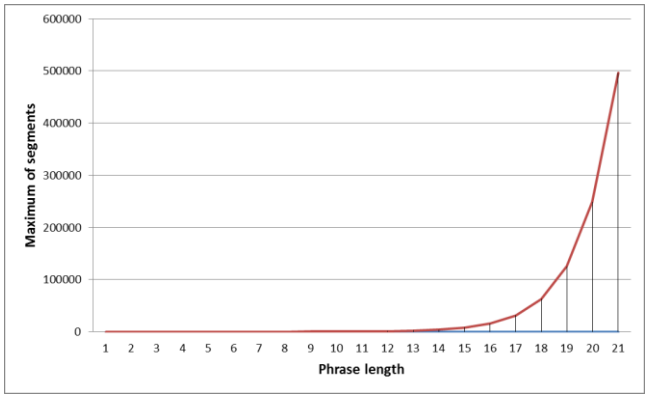

Figure 1. A plot of equation 1 (using a typical maximum word length of 7) showing that the number of possible segments increases at an exponential rate with phrase length.

From Figure 1 it can be seen that the maximum number of segmentations increases exponentially when the phrase length increases, and this leads to the potential risk of "combinatory explosion". Generally speaking, when checking a normal document which has several thousands of words, segmentation alone could cause some delays even when the most of the sentences are normal length and could cause long delays when working on many long sentences [4].

This poster paper appeared at the 4th annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ2013) incorporating the *26h Annual Conference of the National Advisory Committee on Computing Qualifications*, Hamilton, New Zealand, October 6-9, 2013. Mike Lopez and Michael Verhaart, (Eds).

2.2 The Unknown Words

A unique nature of Chinese language is that every character, which is the basic elements in forming a word, has meaning; this is different to the letters in English. This means that new words can be arbitrarily created. Although new words are continually being created in the English language, the effect is much more noticeable in Chinese. Historically, Chinese words mostly consisted of single characters, however more recently the trend has been to create multi-character words that deliver more meaning than single character words. As each Chinese word is built on characters each of which have standalone meaning, it is not practical to define all Chinese words in a dictionary just like it is not practical to pre define all the sentences in English. Another issue is that traditionally, Chinese is a poem like language, especially in writing, and instead of using spaces, people use the rhythm as a natural delimiter. In order to achieve a certain rhythm, it is not unusual that a writer eliminates or adds extra characters in a sentence, and it could be challenge for computer based spell checker to examine these sentences, in other words, a spell checker cannot assume that a sequence of characters that are not in its dictionary are a spelling error. This is one reason that some of the existing Chinese spell checker results in a poor end user experience.

3. A New Approach

Most spell checkers use a list of known correct words to check against, basically all the words in the dictionary. However as already seen, a complete list is impossible to create for a language such as Chinese that continually has new words added and variations for the existing words created. Another approach is to use a list of the most common spelling mistakes. For a language such as English this approach will not be as accurate but for Chinese it has been shown to be a reasonable technique [3]. In fact when Chinese children are learning to write, they are taught a list of around 2000 common spelling mistakes – as experienced by one of the authors. As mentioned above, in the Chinese language, each character has its own meaning and Chinese words are the combination of a variable number of characters, hence, people only make certain kinds of wrong combinations. According to a study [1] there are four main kinds of errors:

1. Misuses of characters due to the same or similar sounds
2. Misuses of characters due to similar shapes;
3. Misuses of characters due to similar meanings
4. Typing errors related to Chinese input methods.

4. THE SOLUTION

4.1 Segmentation

To circumvent the problem of segmentation we have developed the First Character Identifier Approach (FCIA) method. In the first step, the text is scanned one character at a time starting with the 'pointer' pointing at the first character. This character is used to lookup an ordered list of words indexed by their first character. It is possible that the list contains multiple words starting with the same character and this 'sub list' of one or more words is used in the next step. In the second step, the first word in the sub list is used to determine if it matches with the characters at the current position of the text (i.e. starting at the key character). For example given the text AFCSDEAEGKRN and the pointer pointing at the first character A, a sub list of {ABHC, AFC, ACE} might be returned. The word, AFC matches the text. In the third step, the text is now segmented up to the end of the matching word. So now we have the text AFC SDEAEGFRN.

4.2 Combining FCIA with list of known spelling errors

In this work we use a list of incorrect words. If in the previous section a match is made, i.e. a known misspelling is found, the user is then given a choice (from the database) of correct words to choose from. The incorrect word is then replaced by the user's selection or the user can choose to override the system causing the suggested misspelt word to be removed from the list. This removal is done on a per user basis. Note that the misspelt word will be replaced by a word with the same number of characters.

4.3 Moving the Pointer On

If as just described a misspelt word was identified then once the word has been replaced, the pointer is moved on to the next character after the word, S, in this example and the process repeats from this position. If no misspelt word was found, then the pointer just increments one position, to F in this example and the process repeats from there.

5. DESIGN AND IMPLEMENTATION

The algorithm described in section 4 was successfully implemented into the children's email application called Mifrenz (Hunt, 2008) which can be downloaded from http://www.mifrenz.com. Mifrenz implements object orientated and 3 tier design patterns where the classes for: 1) the GUI, 2) the logic and 3) the data storage, input and output, are kept in separate packages. This work followed the same design patterns. Also a number of Use Cases were developed to capture different user scenarios. In order to ensure that the response time for the user was kept to a minimum, the list of words were loaded into a normal Map programming structure as supplied by the Java language instead of using a conventional database.

6. REFERENCES

1 Chang, C. H. (1994). A pilot study on automatic Chinese spelling error correction. Journal of Chinese language and computing, 143-149.

2 Hunt, T. D. (2008). Mifrenz: Safe email for Children. Journal of Applied Computing and Information Technology, 39-51.

3 Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X., et al. (2012). A rule based Chinese spelling and grammar detection system utility. International Conference on System Science and Engineering (ICSSE), (pp. 437-440).

4 Lee, K. H., & Ng, M. K. (1999). Text Segmentation for Chinese Spell Checking. Journal of the American Society for Information Science, 50(9), 751-759.

5 Lin, M. Y., Chiang, T. H., & Sui, K. Y. (1993). A preliminary study on unknown word problem in chinese word segmentation. Proceedings of Rocling, VI, 119-137.