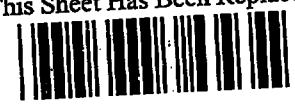
This Sheet Has Been Replaced



ABSTRACT OF THE DISCLOSURE

CHINESE WORD SEGMENTATION APPARATUS

The Chinese word segmentation apparatus of this invention relates to a technique for word segmentation processing of a Chinese sentence inputted into a computer by using character phonetic information in the computer system. A character-to-phonetic converting portion of the Chinese word segmentation apparatus initially converts a Chinese sentence inputted from an input portion of the computer system into a phonetic symbol string while referring to a character phonetic dictionary and a dictionary for characters with different pronunciations. Thereafter, a candidate word-selecting portion refers to a system dictionary to retrieve all of the possible candidate characters or words in the phonetic symbol string and the relevant information, such as frequency of use, etc., using the phonetic symbols as indexing terms. Unfeasible candidate characters or words are discarded via matching means while referring to the characters in the input sentence and syntax constraints of connected candidate words. Subsequently, an optimum candidate character string-deciding portion builds a candidate word network using starting and ending positions of each candidate character or word in the input sentence as indexing terms. By referring to a semantic information portion and a syntax information portion, frequency of use prioritization, word length

prioritization, semantic similarity prioritization and syntax prioritization are combined to obtain a total estimate. The optimum route for word segmentation is then found by a dynamic programming method. Finally, a word segmentation marking portion adds word segmentation markers into the input sentence while referring to the optimum route to complete word segmentation in the Chinese word segmentation apparatus. The apparatus of this invention can achieve a word segmentation accuracy of more than 98%. The invention does not require troublesome and iterative calculations to dramatically increase the operating efficiency and accuracy during Chinese word segmentation.

Fig. 1