CHINESE INPUT SYSTEM WITH ARTIFICIAL INTELLIGENCE

Feng Qian

Computer-Aided Instruction Laboratories, Institute of Educational Technology, East China Normal Univ., Shanghai, China

There are already a variety of Chinese Language Processing Systems commercially available around the world, of which the main divergence is probably in the input approaches of Chinese characters. But due to the following two facts none of the current approaches could be taken for an universally acknowledged scheme, the first fact is that the total number of Chinese characters is immense and the second and perhaps more critical fact is the topological structures of Chinese characters are rather sophisticated.

Considering that Chinese Phonetic Alphabet (CPA) is gaining on Chinese pupils and students (including foreign students studying in China), Considering also that CPA is supposed to be used as an international transcription for Chinese characters, the author has proposed a new input approach to input Chinese language to a computer in the light of AI theory and practice with its prototype just being implement on a microcomputer system in our laboratories. This new Chinese Input System will serve as one of the subsystems of a comprehensive Chinese Language Computer-Aided Instruction System for teaching foreign students on our campus - the HUAHAN system.

Working with our Chinese Input System people can input Chinese text by typing the corresponding phonetic symbols of each Chinese word through an ordinary keyboard. Eventually the original Chinese text will be obtained on the Chinese CRT screen or on the harcopy through Chinese printer.

The bottleneck of this kind of approaches is generally attributed to the large and inreasonable amount of homophones in Chinese language. As a matter of fact, a single Chinese word may consist of more than one characters. In CPA a multi--character word is often represented as one unit. For instance, in phonetic system the string "xingshi" is a word which is formed out of two Chinese characters, i.e. the character "xing" and the character "shi". Some individual Chinese characters, such as "xing" and "shi", may also be a word, hence they may have one or more homophones. Furthermore, multi-character word is still subject to homophone, i.e. phonetic symbol "xingshi" represents both the Chinese words "situation" and "form", among others, though the number of homophones is much reduced.

The key to the question is the software of this Chinese Input System which must be developed as to identify different Chinese words properly on the basis of the same phonetic symbol, i.e., different string of Chinese characters must be generated from the same phonetic symbol occuring in different contexts.

We argue that the differentiation of the homophones could be realized in the similar as in the disambiguation of the same word occuring in different contexts in Natural Language Understanding (NLU) which is making rapid progress.

When considered as without any connection with other words a Chinese word with one or more homophones sharing the same phonetic symbol is really a trouble, but when we try to grasp the proper word not merely by itself but in connection with its context with the background knowledgy and/or with the very topics of the whole text or corresponding paragraph, we find, as a rule, the phonetic ambiguity (i.e. the different homophones which cause language ambiguities) would dissolve. So the most important is to extract the above-mentioned linguistic conceptuations as Chinese text input is going on.

To this end our system is divided into four subsystems including the following:

Lexical analyzer
Context analyzer
Inference mechanism
Knowledge/concept/topic base

The lexical analyzer separates the words which have one or more mohophones from those without homophones.

The context analyzer tries to extract the contextual meaning and/or the topic of the text or the paragraph wherein the word occurs by analyzing the context.

The inference mechanism draws, when necessary, inferences from the contextual meaning in order to obtain proper concept of the troublesome phonetic symbols so as to get the proper Chinese words consequently.

The knowledge/concept/topic base performs as a driver for the whole system, by communicating with each of the other three subsystems in gathering processed results from one subsystem as input data to the other and providing them with additional material necessary for further pocessing.

Thus people can easily see that what our system really does is essentially a Chinese Language Understanding System tries to do. But what our system features as compared with other NLU systems is that we tried hard to develop it to make sure that the knowledge/concept/topic base and the three other subsystems operate concurrently, reflecting our notion of the actual process of human language understanding.

Conventionally Chinese Language Processing is involved in the area of Computer Science. In working up our system we have benefited a lot from theories and practices in Computational Linguistics and advanced researches in AI, especially those research activities at University of Pennsylvania (under direction of Prof. A.K.Joshi) and at Yale University (under direction of Prof. R.C.Schank), all of which bear a strong linguistic flavor.

As a result, the author is inclined to suggest a new interdisciplinary research field be put forward which is tentatively termed as Linguistic Engineering. Our work is taken as a humble start of its practice.

Now the implementation of the system is just going on with a prototype. To our special purpose the BASIC dialects of our CROMENCO and CESEC microcomputers are extended by the author in order to involve some LISP features which would meet the needs to implement such a heavy linguistic system on a microcomputer.