AUXILIARY PROCESS FOR AUTOMATIC ERROR-CORRECTION DURING RECOGNITION OF SOUTH-EAST ASIAN SPEECHES

Koungurtsev Alexey Borisovich, Trinh Nguyen Thieu and Singhal Bijoy Raj.
Faculty of Automation and Computing Techniques
Odessa State Polytechnic University
Address: Odessa-270044, pr. Shevchenko 1, Odessa State Polytechnic University, LINS laboratory, 303 (admin.), Ukraine.
E-mail: thieu@vm.tm.odessa.ua, thieu@lins.ospu.odessa.ua, bijoy_singhal@hotmail.com

ABSTRACT

A speech recogniser converts a spoken sentences into a vector of morphemes and then generates an output sentence in text form. For South-East Asian Languages, unlike the European languages some certain complexities while generating text from the vector of morphemes arise. These complexities are due to intonation, disturbances and differences in voice. As a result the vector of morphemes either contains arbitrary symbols (morphemes which the machine cannot determine) or variations (incorrect morphemes determined by the machine). South-East Asian Languages are monosyllabic in nature i.e. every syllable represents a morpheme. A multitude of these morphemes forms a word. These facts have to be considered while correcting errors during the recognition of South-East Asian speeches. This article presents the idea of an auxiliary process for automatic error-correction while generating the output sentence. The idea is based upon the concepts of word-formation and subject-area.

1. INTRODUCTION

The main tasks accomplished by the Speech Recogniser (SR) are:

- *Spectral analysis*: The output of this process is a set of Sentence-Forming Hypotheses (SFH). Each SFH represents a sentence, which can be formed from the spoken morphemes. Every sentence formed may contain undetermined morphemes and/or a set of similar sounding morphemes.
- *Error-Correction*: At this stage the set of SFH is analysed and one output sentence is generated. This sentence is the actual output from the SR.

At this point, the introduction of an auxiliary process for automatic error-correction, before generating the output sentence, is suggested. The auxiliary process for automatic error-correction includes firstly a process for retrieving undetermined syllables and secondly a process for selecting the most suitable sentence from the possible SFH.

Process for the retrieval of undetermined syllables: For retrieving an undetermined syllable it is necessary to consider the possibilities of its forming a word with the morphemes located before and after the given syllable. After the completion of this process a set of SFH, free of any undetermined syllables, is obtained.

Process for selecting the most suitable sentence: For the elimination of useless SFH from the set of all possible SFH, it is necessary to consider morphological and semantic validity of words.

The main obstacle in the determination of morphological validity arises due to the difficulty in determining limits of a word in a sentence [2][3]. It is necessary to check all possibilities of word-formation in a sentence to overcome this obstacle. The traditional method (enumeration method) used to check these possible word-formations requires considerable machine time resources. So some other methods for determining word-formations are suggested. Their effectiveness is compared and an optimal algorithm is represented. Apart from this, a method for eliminating the useless SFH based on the concept of subject-area is suggested. This reduces the processing time considerably.

2. REPRESENTATION OF THE PROBLEM

Representation of problem: “*Determine morphological and semantic correct SFH from the set of all SFH*”.

For this, it is necessary to solve the following sub-problems:

Sub-problem 1: “*Determine all words from the vector of morphemes representing sentence S*”.

Sub-problem 2: “*Determine a semantic correct SFH from the set of SFH*”.

To solve the sub-problem 1, it is necessary to solve the following problem: “*Determine all words from the vector of morphemes representing fragment P”*.

Let us consider the possible solutions to the problems.

3. METHOD FOR CORRECTING ERRORS CONSIDERING MORPHOLOGICAL VALIDITY

In South-East Asian Languages (SEAL) words are formed from one or many morphemes. Let us denote:

- set of words:

$\text{Word} = \{W_i\}, \forall i \in N,$ where W_i — word, N — natural number;

- set of morphemes:

$\text{Morph} = \{e_i\}, \forall i \in N,$ where e_i — morpheme,

Let us call:

- P as a fragment of sentence. Every fragment consists of a series of words and is separated from the other fragments by pauses, i.e. $P = [\{W_i\}], \forall i \in N;$
- Sentence as a vector of fragments and pauses, i.e. $S = \{\{P\}\}.$

In a sentence a word is defined as a chain of morphemes, i.e. $W_i = [\{e_i\}], \forall e_i \in \text{Morph}.$

3.1. Enumeration method

Given $P = [\{e_i\}] = [e_1, e_2, \ldots, e_n], \forall n \geq 1, e_i \in \text{Morph}.$

P is a fragment of a sentence if and only if a vector of words:

$$[W_1, W_2, \ldots, W_m],$$
$$W_j = [e_i, e_{i+1}, \ldots, e_{i+k}],$$
$$W_{j+1} = [e_{i+k+1}, e_{i+k+2}, \ldots, e_{i+k+x}],$$
$$1 \leq m \leq n, 1 \leq j \leq m,\ 1 \leq i \leq n,\ k \geq 0,\ x \geq 1,$$

can be formed from vector $[e_1, e_2, \ldots, e_n]$.

The essence of the enumeration method lies in checking whether the vector $[e_i, e_{i+1}, \ldots, e_{i+k}]$ is a word or not. The solution is a combination of all possible alternatives of the fragment-formation hypotheses. It can be proved that 2^{n-1} checks are necessary to determine all possible fragment-formation hypotheses.

3.2. Method based on vectors of word-formation possibilities (L-method)

Let us denote $e_{i:i+1:\ldots:k}$ as a word formed by series of morphemes from e_i to e_k. During analysis of fragment P, using the enumeration method, it is necessary to check the following combinations:

$$K_{k+1} = \{[e_1, e_2, \ldots, e_{n-1}]e_n, [e_1, e_2, \ldots, e_{n-2}]e_{n-1:n}, \ldots,$$
$$[e_1, e_2, \ldots, e_i]e_{i+1:i+2:\ldots:n}, [e_1, e_2, \ldots, e_{i-1}]e_{i:i+1:\ldots:n}, \ldots,$$
$$[e_1]e_{2:3:\ldots:n-1}, e_{1:2:\ldots:n}\}.$$

It is clear that while verifying combinations $[e_1, e_2, \ldots, e_i]e_{i+1:i+2:\ldots:n}$ verification of combinations $[e_1, e_2, \ldots, e_{i-1}]e_{i:i+1:\ldots:n}$ is repeated. For eliminating the repetition it is necessary to pre-determine all possible word-formations based on morphemes of the given fragment P. For this let us introduce the term — set of possible combinations:

$$L = [\{L_i\}],\ 1 \leq i \leq n,$$
$$L_i = [\{l_{ij}\}],\ 1 \leq i \leq n,\ 1 \leq j \leq n - i + 1,\ l_{ij} \in \{0;1\}.$$

From vector L we can determine whether or not the combination $e_{i:i+1:\ldots:i+j}$ is a word. If $l_{ij}=1$, then — YES, otherwise — NO. In L_i, all possible word-formations starting from i-*th* morpheme are stored. The verification carries on from e_1 to e_n. For every e_i, it is necessary to check another n-i+1 combinations of $e_{i:i+1:\ldots:j}, (i \leq j \leq n - i).$ So the set L consists of *n* vectors L_i and each vector L_i is the size of n-i+1. Hence, the number of checks is:

$$\sum_{i=1}^{n}(n - i + 1) = \sum_{i=1}^{n}(n + 1) - \sum_{i=1}^{n} i = n(n + 1)/2$$

Based on the set L, we can determine whether the given fragment is correct or not. For this we convert the set L into set U:

$$U = [\{U_i\}],\ U_i = [\{u_{ij}\}],\ 1 \leq i, j \leq n,$$
$$u_{ij} = \begin{vmatrix} 0, & (1 < i < j), \\ l_{i(j-i+1)}, & (i \leq j \leq n - i + 1). \end{vmatrix}$$

The following indication has been proved:

Indication 1: Fragment P is incorrect when:

$$\exists i,\ 1 \leq i \leq n,\ u_{ij} = 0, \forall j,\ i \leq j \leq n.$$

This indication implies:

Consequence 1: Fragment P is incorrect when:

$$u_{1j} = 0, \forall j,\ 1 \leq j \leq n.$$

Consequence 2: Fragment P is incorrect when:

$$u_{in} = 0,\ \forall i,\ 1 \leq i \leq n.$$

Realization of L-method consists of the following steps:

1. *Step 1*: Verify all possible combinations of words based on morpheme e_i, $(\forall i,\ 0 \leq i \leq n)$ in the given fragment P. The results are written into the vector L_i.
2. *Step 2*: Checking the lexical correctness of fragment P by indication 1 and its consequences. In case of negative result jump to step 4.
3. *Step 3*: Generate solutions based on set U in ascending order. For this a tree of solutions is constructed using the following rules:
   - Create empty root G,
   - Create connected nodes G_{in} by condition:

     $\text{for } \forall i,\ 1 \leq i \leq n,\ \text{if } u_{in} = 1, \text{then } G = [\{G_{in}\}],$
   - For every node G_{in} recursively construct connected nodes G_{jk} by condition:

     $\text{for } \forall j,\ 1 \leq j \leq k, k = i - 1,\ \text{if } u_{jk} = 1,$

     $\text{then } G_{in} = [\{G_{jk}\}],$
   - For j=1, the general form of solution is: $e_{1:2:\ldots:p}\, e_{p+1:\ldots:x} \ldots e_{q:q+1:\ldots:n}$.
4. *Step 4*: End.

3.3. Method based on the number of morphemes in a word (Q-method)

Let us denote *maxw* — maximum number of morphemes forming a word in the electronic dictionary. It can be easily proved that $l_{ij} = 0, \forall j > \text{maxw}.$ Therefore, for the formation of vector L_i, it is reasonable to verify combinations only from e_i to $e_{i+maxw-1}$. It can be seen that for finding all possible fragment formation hypotheses it is necessary to conduct Q checks. It can be proved that:

$$Q = \begin{cases} \dfrac{n(n+1)}{2}, & (n \le \text{maxw}), \\ \dfrac{\text{maxw}(2n - \text{maxw} + 1)}{2}, & (n > \text{maxw}). \end{cases}$$

In monosyllabic languages, majority of the words consist of two morphemes [3]. Words formed from three or four morphemes have the following properties:

Let us see $W = [e_1, e_2, \ldots, e_m]$, $(2 < m \le 4)$, then:

- Either the morpheme e_j, or the combination $e_{j:j+i}$ $(1 \le i \le m)$ is not a word;
- Either the morpheme e_j, or the combination $e_{j:j+i}$ $(1 \le i \le m)$ is a word, but fragment P is grammatical or semantic incorrect.

So it is reasonable to conduct checks only for words consisting up to two morphemes (i.e. only combinations e_i and $e_{i:i+1}$). If the expected result is not found then the remaining checks of up to four morphemes are conducted.

3.4. Optimal algorithm for determination of word-formations

It can be proved that benefits of the L-method and Q-method with respect to the enumeration method can be determined using the following relations:

$$h_L = \frac{T_{enum}}{T_L} = \frac{2^n}{n(n+1)}, \quad h_Q = \frac{T_{enum}}{T_Q} = \frac{2^n}{n(\max w + 1)},$$

where T_{enum} — time of enumeration method, T_L — time of L-method, T_Q — time of Q-method.

It is easy to prove that L-method will be effective when $n \ge 5$ and for Q-method when $n > 3$. For instance for maxw = 4, when n = 10 , h_L = 9.3, h_Q = 20.48; when n = 20, h_L=2496.9, sh_Q = 10458.76.

Based on the analysis of specialities of word formation in SEAL, the following algorithm for formation of words in a sentence was obtained:

1. *Step 1*: Determination of fragments during spectral analysis. It is accepted that $2 \le \text{maxw} \le 4$.
2. *Step 2*: Searching all possible words for every fragment. The search method is selected based on the condition:
    - 2^n - n(maxw+1) ≤ 0 — enumeration method;
    - 2^n - n(maxw+1) ≥ 0 — Q-method.
3. *Step 3*: Generating solution. In case of its absence increase maxw and jump to step 2.
4. *Step 4*: Completion of syntactic analysis and semantic analysis. If there is no solution increase maxw, and jump to step 2.
5. *Step 5*: End.

4. METHOD FOR CORRECTING ERRORS CONSIDERING SEMANTIC VALIDITY

After the correction of morphological errors, there is a possibility that some SFH, which do not have a semantic validity, are left. For selecting the correct sentence, a selection method based on the concept of subject-area is suggested. The subject-area means some constants denoting semantics of the word. So the SFH is considered semantic correct if the following condition is satisfied:

$$\text{Subj\_ar}(w_1) \cap \text{Subj\_ar}(w_2) \cap \ldots \cap \text{Subj\_ar}(w_m) \ne \varnothing,$$

where: Subj_ar(w_i) is the subject-area of the word w_i in SFH. This means that all words belong to the same main subject-area.

Method for correcting errors considering the semantic validity allows the effective determination of the required solutions. This permits the SR to determine the mistakes of the speaker [1]. The drawback of this method is that, it requires a separate knowledge base containing a hierarchy of subject-areas.

5. CONCLUSION

The use of vector L, for the analysis of the word formation specialities of SEAL, permits us to economise analysis time. The verification of word-formation possibilities takes place parallel to the input process from microphone. After the sentence is spoken, the analyser concludes the process of L-vector formation. The rest of steps don’t require considerable machine time resources. Besides, the mechanism for the auxiliary process can be used for text-understanding also.

6. REFERENCES

[1] Alieva N.F. Article collection of The Scientific Academy of USSR, Institute of Eastern Studies. 1980. *South-east Asian languages: Problems of word repeating.* Moscow: Nauka.

[2] Omelianovich N.V. Article collection of The Scientific Academy of USSR, Institute of Eastern Studies. 1985. *South-East Asian languages: Problems of complex words.* Moscow: Nauka.

[3] Soltseva H.V. Article collection of The Scientific Academy of USSR, Institute of Eastern Studies. 1970. *South-East Asian languages: Problems of morphology, phonetics and phonology*. Moscow: Nauka.