WORD BOUNDARY HYPOTHESISATION IN HINDI SPEECH

G.V.Ramana Rao*, M.PraKash** and B.Yegnanarayana*

* Dept. of Computer Science and Engg, I.I.T., Madras, India
**Dept. of Electronics and Computer Science, Regional Engineering College, Warangal, India

ABSTRACT

In this paper we report the work done on the hypothesisation of word boundaries in Hindi speech. Firstly, the language related clues useful in word boundary hypothesisation are identified. These clues are then grouped into two classes: (i) clues describing conditions to be satisfied at word boundaries, called lexical clues and (ii) clues that consist of frequently occurring patterns in text, called pattern clues. The measure of frequency and correctness, are used to evaluate these clues. The results led to the organization of these into hypothesisation and verification clues. Experiments with the word boundary hypothesiser on a 300 sentence text yielded two results: (i) upto 50% of word boundaries in the text are hypothesised and (ii) considerable speed improvement in lexical search is achieved due to the hypothesisation of these boundaries.

1 INTRODUCTION

Continuous speech offers few clues for hypothesising word boundaries. Hence current speech recognition systems perform word boundary and word hypothesisation together. This usually is a dictionary search that tries to match the input symbol sequence (the output of acoustic-phonetic analysis) with the prestored lexicon. However, reliable hypothesisation of word boundaries greatly helps in reducing this lexical search. In a few cases, prosodic features such as pause and duration can be used to hypothesise some of the word boundaries. However such clues are too few to be of real help. The work reported in this paper focuses on the use of language dependent clues in word boundary hypothesisation.

The word boundary hypothesisation problem can be posed as follows. Given a sequence of symbols representing an utterance, what clues are available in the language to place the word boundaries. Posed in this fashion, the word boundary hypothesisation becomes dictionary independent and hence it can be separated from the problem of word hypothesisation. Thus one could build a word boundary hypothesiser and use it as a preprocessor to the word hypothesiser. The advantages of this approach over normal word hypothesisation are as follows:

(i) If most of the word boundaries could be hypothesised, an inexpensive speech recognition system could be built.
(ii) The lexical search in the word hypothesisation process could be speeded up.
(iii) Since word boundary hypothesisation partitions the utterance into smaller subsequences of words the lexical search can be performed in parallel over each of the subsequences.
(iv) It offers an elegant way of handling unknown words.

In a typical speech recognition system, an acoustic-phonetic analyzer produces a sequence of symbols representing the utterance. Higher level analyses such as lexical, syntax, and semantics are performed on this symbolic representation to produce the final text. However, a speech recognition system has a human as an end user (unlike an understanding system) and hence depending on the task, one can eliminate the higher levels of processing which are quite complex to build and thereby gain in system speed and simplicity at the cost of accuracy. Word boundary hypothesisation fits well in this approach, since acceptable quality of text can be produced if one has a good acoustic-phonetic analyzer followed by a word boundary hypothesiser. This is illustrated by the example shown in Fig.1. Fig.1(a) shows a sentence with the word boundaries removed. Figures 1(b), 1(c) and 1(d) show the same sentence with 30, 70 and 100 percent of the word boundaries put in. From these one could clearly see the improvement in recognition as more and more boundaries are placed.

Another important result of word boundary hypothesisation is the reduction in the lexical search (word hypothesisation) complexity. It can be noted that this complexity usually varies exponentially with respect to the length of the input symbol sequence. Word boundary hypothesisation simplifies the lexical search since it breaks the symbol sequence into smaller subsequences. Further improvement in the lexical search speed is possible due to the fact that one can now conduct the lexical search over the various subsequences in parallel.

The ability to hypothesise word boundaries independent of the dictionary also simplifies the handling of unknown words, i.e., words not present in the dictionary, such as proper nouns. In a
normal word hypothesiser, an unknown word is hypothesised when the lexical search returns a 'no match'. In such cases the recovery is quite cumbersome since one does not know the beginning of the next word to continue the search. Hence one has to skip each symbol of the unknown word and try for a dictionary match from the next symbol until the beginning of the next word is reached, resulting in many unnecessary dictionary searches. However, in an utterance where some of the word boundaries are already hypothesised, one could use the simple recovery strategy to jump to the next word boundary and continue the search from there.

In Section 2, we discuss the language related clues useful for word boundary hypothesisation in Hindi speech. A word boundary hypothesiser based on these clues is developed and its performance is presented in Section 3.

2 CLUES FOR WORD BOUNDARY HYPOTHESISATION

In this section we describe the clues useful in identifying word boundaries. Only language related (LR) clues are considered for this purpose. Earlier work by several linguists, notably [1] and [2], identified several clues for word boundary hypothesisation in Hindi speech. A list of these is given in Fig.2. As these clues describe the conditions to be satisfied at word boundaries, these are named 'lexical clues'. Unfortunately these lexical clues are few in number and as shown in later sections are not very effective in hypothesising word boundaries. Therefore other clues based on the word frequency analysis have been identified. These clues are based on the fact that different words have different frequencies of usage and hence if one looks for patterns corresponding to the more frequently used words, then the probability that the pattern found is actually the hypothesised word will be quite high. Thus the nonlexical clues are nothing but frequently occurring patterns (not necessarily words) are also included among these clues. Hence the word boundary hypothesisation using these nonlexical clues (henceforth referred as 'pattern clues') is equivalent to the spotting of certain common patterns in the input symbol sequence.

Two factors are primarily considered in selecting these common patterns. First is that they should occur frequently and secondly, they should be quite general, i.e., they should occur in all types of text. Under these criteria, only case markers and certain other classes like pronouns and conjunctions qualified as pattern clues. In addition, certain verb endings are also included in the pattern clues. These patterns are small in number (approximately 25) and they serve important syntactic functions. The case markers are related to noun phrases and hence they can be useful in parsing of phrases. A parsing approach based on the spotting of the case markers is under development [3]. Similarly verb endings and conjunctions serve as clause markers and hence can be used in the parsing of clauses.

3 WORD BOUNDARY HYPOTHESISATION

Three studies are conducted on Hindi speech using the above discussed clues for word boundary hypothesisation. The first study measures the applicability and the correctness of each clue in hypothesising word boundaries. The second study is on the development of a word boundary hypothesiser for Hindi and finally the third study measures the improvement in lexical search due to prior word boundary hypothesisation. Each of these studies is detailed below.

3.1 Evaluation of Word Boundary Hypothesisation clues

In this study the word boundary hypothesis clues are evaluated using two measures, namely, frequency and correctness. The frequency measure is used to indicate how useful a clue is in terms of the number of word boundary hypotheses it produces. It is defined as the ratio of the number of times a word boundary is correctly hypothesised to the number of word boundaries actually present in the input. The correctness measure indicates the confidence in the hypotheses made using the clue. It is defined as the ratio of the number of times a word boundary is correctly hypothesised to the number of times a word boundary is hypothesised using the clue.

The clues are evaluated using a 1000 sentence text containing nearly 10000 word boundaries. All the word boundaries are removed from this text and word boundaries are hypothesised using each clue. From these, the frequency and correctness measures are calculated for each of the clues. These results are shown in Fig. 3. The results for the pattern clues are shown in three groups namely, case markers, verb endings and pronouns and conjunctions (shown in Fig.3. as 'other keywords').

3.2 Development of Word Boundary Hypothesiser

As the correctness and frequency measures of the lexical clues show, their utility in hypothesising word boundaries is rather low. For example, the lexical clue LR1 in Fig. 3. has a very high frequency value but its correctness is so low that more erroneous than correct hypotheses will be generated if it is used to hypothesise the word boundaries. On the other hand, lexical clues LR2, LR3 and LR4 have very high correctness but their frequency is low. Hence their utility in producing word boundary hypotheses is negligible. However, pattern clues,
especially case markers and pronouns and conjunctions have reasonably good frequency and correctness values. Hence these could serve as good clues for word boundary hypothesisation.

Based on the above, the word boundary hypothesiser is organized as follows. The pattern clues are used to hypothesise word boundaries. This in practice is done by simple pattern matching. The patterns corresponding to the pattern clues are organized as a TRIE and it is matched against the input sentence. At the boundaries hypothesised the lexical clues are applied as verification rules. Since these clues have a high correctness (LR1 is applied in a verifying sense) they help in weeding out some of the erroneous hypotheses generated by the pattern clues. The corresponding results for word boundary hypothesisation are shown in Fig. 4.

While the above results demonstrate the validity of our word boundary hypothesiser, from the point of improvement in lexical search a more relevant result is the size of the sub-sentences after the word boundary hypothesisation. If the word boundary hypothesiser still leaves a large number of long subsentences then there could hardly be any improvement in the lexical search speed. In this regard a plot of the distribution of the subsentences (obtained after the word boundaries are placed) with respect to the size is shown in Fig. 5.

3.3 Improvement in Lexical Search

In this study the relative improvement in lexical search speed is measured. A simple lexical analyzer is built to hypothesise a sentence from the input symbol sequence in a best first fashion. A dictionary of 1500 words is used for this purpose. The dictionary is organized as a TRIE. A sample text of 334 sentences (from 6 stories) is used as input to the lexical analyser. The time taken to produce the sentence hypotheses is measured for both normal text and for text output by the word boundary hypothesiser. The results are shown in Fig. 6.

The results of the above studies show that significant number of word boundaries can be hypothesised using these clues. Also considerable speed improvement in lexical search is also observed. If parallel search is also employed, further reduction in the lexical search time can be achieved. However, certain points are to be explained before the conclusion:

(i) The number of word boundaries that can be hypothesised are limited: As the results show the number of the correct word boundaries that are hypothesised seem to be around 50 percent at most. This is due to the fact that many sentences contain phrases that do not contain any of the patterns used for hypothesising boundaries. For example, a noun phrase of the form 'adjective(s) followed by a noun' can be further broken by our approach. Secondly, a number of function words are not included in the pattern clues because of their low frequencies of occurrence. This also resulted in missing some of the word boundaries.

(ii) The lexical search speed improvement is smaller than expected: Though significant speed improvement (approx. 40%) is observed, it is still smaller than expected because many of the input sentences are small (less than 10 words). Since in small sentences the chance of hypothesising a boundary is low, there will not be much speed improvement in these sentences. In fact, the improvement in speed for sentences of size larger than 10 words is much higher (around 60%). Another factor that reduced the lexical search speed is the erroneous boundary hypotheses which are around 12 percent.

4 CONCLUSIONS

The work presented here resulted in the identification of some language dependent clues for hypothesising word boundaries in Hindi speech. These clues are evaluated with respect to their utility in word boundary hypothesisation. The improvement in lexical search speed in the presence of word boundaries is also estimated. The results demonstrated the utility of word boundary hypothesisation in speech recognition.

REFERENCES

1. Kailash Chandra Bhatia, हिंदी भाषा में अंग्रेज़ तथा अन्य भाषाओं की सीमाएँ
(Syllable and Word Boundaries in Hindi), Nagari Pracharini Sabha, Varanasi, 1970.

(a) उसकी साथ जो उसकी मित्रता करती जो रायत मतलब से दूर दूर रहती
(b) उस की साथ वह जो रायत करता जो रायत करता भारती धर्म से दूर दूर रहती
(c) उस की साथ वह जो रायत करता जो रायत करता धर्म से परतु वह उस से दूर दूर रहती
(d) उस की साथ वह जो रिता जो रायत करता धर्म से परतु वह उस से दूर दूर रहती

Fig. 1. Illustration of use of word boundaries in human recognition.
(a) A Hindi sentence with all the word boundaries removed.
(b) The sentence with a few boundaries (30%) inserted.
(c) The sentence with more boundaries (70%) inserted.
(d) The sentence with all the boundaries inserted.
LR1: A Hindi word can end in either a long vowel or a consonant except for the words 'na' (ना) and 'ki' (की).

LR2: Only certain consonant sequences can occur at word initial position.

LR3: Only certain consonant sequences can occur at word final position.

LR4: Only certain vowel sequences can occur in word initial position.

Fig.2. A list of the lexical rules. (For details refer [1].)

<table>
<thead>
<tr>
<th>Clue</th>
<th>Frequency</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR1</td>
<td>0.99</td>
<td>0.24</td>
</tr>
<tr>
<td>LR2</td>
<td>0.01</td>
<td>0.95</td>
</tr>
<tr>
<td>LR3</td>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>LR4</td>
<td>0.04</td>
<td>0.93</td>
</tr>
<tr>
<td>Case markers</td>
<td>0.26</td>
<td>0.86</td>
</tr>
<tr>
<td>Verb endings</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>Other keywords</td>
<td>0.16</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Fig.3. The frequency and correctness values for the word boundary hypothesisation clues.

Fig.4. Results of word boundary hypothesisation.

<table>
<thead>
<tr>
<th>Clue</th>
<th>Number of Correct Boundaries (% of total)</th>
<th>Number of Erroneous Boundaries (% of hypotheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case markers</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>Verb endings</td>
<td>3.6</td>
<td>10</td>
</tr>
<tr>
<td>Pronouns and Conjunctions</td>
<td>15.4</td>
<td>11</td>
</tr>
<tr>
<td>All Clues Together</td>
<td>44</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Fig.5 (a)

Fig.5 (b)

Fig.5. Distribution of the subsentences (after word boundary hypothesisation) w.r.t. their size.

(a) Plot of the distribution of input sentences.
(b) Plot of distribution of subsentences after word boundary hypothesisation.

<table>
<thead>
<tr>
<th>Time (sec.)</th>
<th>% of Normal Search Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Word Boundaries</td>
<td>30.7</td>
</tr>
<tr>
<td>Word Boundaries Hypothesised From Case Markers</td>
<td>21.5</td>
</tr>
<tr>
<td>Word Boundaries Hypothesised From Verb Endings</td>
<td>30.3</td>
</tr>
<tr>
<td>Word Boundaries Hypothesised From Other Keywords</td>
<td>26.7</td>
</tr>
<tr>
<td>Word Boundaries Hypothesised Using All Clues</td>
<td>18.2</td>
</tr>
</tbody>
</table>

Fig.6. Comparison of lexical search times for sentences with and without word boundaries.

EUROSPEECH '89, Paris, France, September 1989 2363