DETECTION AND CORRECTION OF PHONETIC ERRORS WITH A NEW ORTHOGRAPHIC DICTIONARY

Dr. Sivaji Bandyopadhyay

Computer Science and Engineering Department
Jadavpur University, Calcutta, INDIA.

E-mail : ilidju@cal2.vsnl.net.in

ABSTRACT

The Phonetic and the Homophone Error problem in a language have been characterized as a symbol substitution problem. Phonetically equivalent symbols or symbol combinations in the language are grouped together. Each group or a number of related groups give(s) rise to a dictionary or a number of dictionaries. A new design methodology for Orthographic Dictionaries in alphabetic languages has been described. The dictionaries include the root words. The meanings are stored only in case of Homophone words. Words are sorted on the basis of a Phonetic Ordering Scheme. The dictionaries are being used to detect and correct the Phonetic Error and the Homophone Error in isolated words of Bengali. The methodologies described in this paper can be used in developing spell-checkers not only for other Indian languages which have evolved from the ancient Brahmi script and have a common phonetic structure but also for other alphabetic languages with suitable modifications.

1. The Phonetic Error Problem in NLP - an Introduction

The practical NLP systems must deal with erroneous or ill-formed inputs. The nature of errors basically depends on the source of input as well as on the user's exceeding the systems limited grammatical, conceptual or functional coverage. Automatic word correction research may be viewed [7] as focusing on three increasingly broader problems: (1) nonword error detection; (2) isolated word error correction; and (3)context dependent word correction. Work on the isolated-word error correction spanned from as early as the 1960s into the present. Distinctions are sometimes made among three different types of nonword misspellings : (1) Typographic Errors - the writer or the typist knows the correct spelling but simply makes a motor coordination slip, (2) Cognitive Errors - due to misconception or lack of knowledge on the part of the writer or the typist and (3) Phonetic Errors - a special class of cognitive errors in which the writer substitutes a phonetically correct but orthographically incorrect sequence of letters for the intended words. Moreover, there exists a class of real-word errors in which the misspellings result in a valid word. It occurs due to the presence of words in the language having similar pronunciation but different meaning. This is also termed as Homophone Errors.

With an increase in Natural Language applications using Indian Languages, it has become necessary to study the reasons and the nature of errors that can occur in these languages and how these errors can be

detected and corrected. In the present work, the Phonetic and the Homophone error problem in Bengali have been studied in detail.

2. The Phonetic and the Homophone Error Problem in Bengali

Bengali is an important language in India and has evolved from the ancient Brahmi Script. The people in the neighboring country of Bangladesh use Bengali as their Official Language. The discrepancy between the spelling of a word in Bengali and its pronunciation has been generated due to the reason that the Sanskrit alphabet has been fully taken up while the symbols are pronounced in the Bengali way. The Phonetic Error problem in Bengali can be characterized as a substitution problem [12,13]. The Homophone Error problem can also be looked upon as a substitution problem in which the substitution of a correct symbol in a valid word by its phonetic equivalent generates another valid word in the language.

A vowel may be substituted by another vowel or by a consonant along with a matra or vice versa. A consonant may be substituted by another consonant or a conjunct or vice versa. A conjunct may also be substituted by another conjunct or by a consonant along with a specified matra or by a diacritic attached to the previous unit and a consonant in the present unit or vice versa. A matra may be substituted by another matra or by the absence of a matra.

The problem of detection and correction of spelling errors in English has been investigated by several workers [2]. The minimum distance criterion or statistical properties as used in English are likely to result in in-efficient implementation in case of Hindi and other Indian languages [1].

The work by Sinha & Singh (1984) [1] was done for correction of single spelling errors in Hindi words using Dictionary look - up. The implementation reported by Srinivas and Ravisankar (1990) [3] is based on a heuristic approach using the trie as the basic data structure for representing the lexicon. Ranade and Bhaskararao (1992) [5] developed the identification procedures for input Marathi words with applications for a spelling checker. The work done by Bhatt, Doctor & Bhaskararao (1992) [6] is mainly concerned about the implementation of a spell-checker for written Hindi languages. A spell-checker in the GIST (**G**raphics and **I**ntelligence based **S**cript **T**echnology) has been developed by the Center for Development of Advanced Computing (CDAC), Pune, India. It is known as SPELL CASTER and is currently available for Hindi, Marathi, Gujarati and Tamil.

The detection and correction of spelling errors in the Bengali script are reported by Chaudhuri & Pal (1996,1998) [8,9] and Bandyopadhyay (1998) [10].

3. The System Functions

The Phonetic Error Detection and Correction Software functions at word level. The input can be a valid root word or an inflected word or a word starting with numerals and ending with a letter or a compound word with a hyphen. Some typographic errors are checked during input.

Any phonetic errors present in the word are detected during dictionary search as well as through the use of rule-based techniques. The following outcomes are possible for an input word :

(a) If the word starts with numerals the digits at the start are bypassed and the unit (i.e., the letter) that can be present after the numerals is checked for spelling corrections using appropriate rules.
(b) It can be a root word and there is no possibility of any phonetic error. Such words are not included in the dictionaries.
(c) It can be a root word and there is possibility of phonetic error. If the word is present in one of the dictionaries then either,
(i) the word is correctly spelled, or
(ii) all the phonetic errors in the word are fully corrected.
If the word is not present in any one of the dictionaries then either,
(iii) some phonetic error corrections may be applied on the word, or
(iv) no phonetic error corrections can be applied.

Since, only one candidate correction is generated in (iii), the question of ranking the candidate corrections does not arise.

(d) If the word is found in a dictionary, it may turn out to be a homophone. All the homophones and their meanings are provided as output.
(e) If the input compound word includes a hyphen, the outcome can be one of (b), (c) or (d). The presence of hyphen in a word is ignored during comparison.
(f) If the word can not be checked successfully as a root word, it may be an inflected word. The outcome for the root form can be one of (b), (c) or (d).

In the present work, a new design scheme for Orthographic Dictionaries [10,11] has been developed. The script processing is done using a technique which includes modifications from the current ISCII (Indian Script Code for Information Interchange) standard [4]. The techniques that are developed can be used in designing Spell - Checkers for other Indian Languages which have evolved from the ancient Brahmi Script and have a common phonetic structure.

4. Internal Storage of a Word : The UNIT Concept

An Unique Identification Number is assigned to each symbol in Bengali, i.e., vowels, consonants, conjuncts, matras, diacritics, numerals and special symbols, based on the collating sequence. This Symbol Number is derived from the ISCII codes of each basic symbol. The details of this Symbol Number allocation can be found in [10].

Each letter in Bengali is stored internally as a structure or a unit containing the Symbol Numbers of the character, the matra and the diacritic. The character can be a vowel, a consonant or a conjunct and will always be present. The Hard Halant in a letter is considered as its matra. The absence of the matra and/or the diacritic will be indicated by storing a 000 in the respective position. The Symbol Numbers for the numerals or the special symbols are included in the character part of a unit, the matra and the diacritics are considered as absent. The Space character is represented by storing 000s in all the three positions. This scheme is also applicable for other Indian languages.

The salient features of the unit concept are :

(a) The phonetic ordering of two letters is now based on comparing their constituents in the following order : the character, the matra and the diacritic.
(b) The substitution of a conjunct by the correct consonant or vice versa becomes very straightforward.
(c) The method of storage for numerals and special symbols is useful during editing when such a symbol may have to be replaced by a letter.
(d) Sometimes, a no matra in a unit can be misspelled as a matra. Similarly a no matra may have to be replaced by a matra. Sometimes the character and the matra combination in a word may have to be replaced by a character only. The internal storage requirement of a word remains the same after any possible correction since space is reserved in a word for its three components.
(e) The phonetic errors associated with Visarg and Chandrabindu diacritics are in their omission or addition. Such cases can be effectively dealt with since a placeholder for diacritic is kept.

5. Phonetic Equivalence Groups

Phonetically equivalent symbols in Bengali have been grouped together under different **Phonetic Equivalence Groups**. The equivalence between symbols or symbol combinations in a group depends on several factors like the position of the symbol in a word (start, middle or end) and the matra attached to the symbol. A total of 135 **Phonetic Equivalence Groups** have been identified in Bengali.

A possible phonetic error can be detected in an input word by looking at all the three symbols (character, matra and diacritic) in each and every letter and suspecting the symbols which belong to an equivalence group. Phonetic Error detection and correction procedures are written for these groups. Similarly, orthographic dictionaries have also been designed on the basis of the Phonetic Equivalence Groups. During the comparison of the input word with a word in the dictionary, if the symbols at a particular position do not match but the symbols belong to the same group then an error is signaled. The symbol in the input word is changed to the corresponding symbol in the dictionary word.

The Vowel, the Consonant and the Matra groups are always active. These are termed as the **General Equivalence Groups**. On the other hand, there are some Phonetic Equivalence Groups which are divided into some subgroups on the basis of the matra that may be attached to the symbols of the group. These subgroups are known as the **Special Phonetic Equivalence Groups** and have been assigned a unique **Special Phonetic Equivalence Group Number.** On the basis of the number set by a Phonetic Error Detection and Correction procedure, the appropriate subgroup is taken into consideration. If a number of symbols are present in a word which belong to Equivalence Groups other than the General Equivalence Groups then more than one procedure may be executed. All the Phonetic Error Detection and Correction procedures will work on the word but it will be stored in a single dictionary. Thus, distribution of words into the various dictionaries is crucial to the functioning of the system.

6. The Orthographic Dictionaries

Each Phonetic Equivalence Group or a number of related groups give(s) rise to a dictionary and in some cases to a number of dictionaries. If a word contains only one symbol belonging to one of the Phonetic Equivalence Groups then it will be included in the corresponding dictionary. Methodologies have been

developed to distinctly identify the dictionary when a word has more than one symbol belonging to the same or different Equivalence Groups. The dictionaries are orthographic in nature as only the correct spellings of the words are included and there are no associated semantic and / or grammatical information. The meanings of words are included only in the case of Homophone words in order to select the appropriate word in the current context.

Each dictionary is stored as a flat file and each word is of fixed length. The Homophone Area of a dictionary is placed just after the Normal Word Area. All the Homophones of a word are given the same Homophone Number. In some dictionaries, all the Homophones of a word may be present consecutively in the Normal Word Area also if they satisfy the criteria for inclusion. Otherwise, some or at least one Homophone of the word are (is) also stored in the Normal Word Area.

The words in a dictionary are organized into several Index groups. The words within an Index group are ordered phonetically. The **Phonetic Ordering Scheme** is basically similar to the alphabetic ordering scheme except that symbols belonging to the same Phonetic Equivalence Group have the same phonetic order. Whenever, two corresponding units mismatch in the input word and in the dictionary word but there is equivalence either in the character or in the matra then the input word symbol can be corrected to that in the dictionary word. All the misspelled symbols in the input word would be corrected if the correct spelling for the word is present in one of the dictionaries. A correctly spelled word will be directly obtained in the dictionary.

Search procedures have been written to search a word in a dictionary. Each successive word under an appropriate Index Record is read and compared with the input word. During the comparison operation, any possible corrections are attempted. The comparison is carried out on a unit basis and continues until one of the words is fully scanned.

7. The Phonetic Error Detection and Correction Procedures

The Control procedure for monitoring the overall process of Phonetic Error Detection and Correction initially checks whether the input word starts with numerals. If it is so, the validity of the word is checked. Otherwise, each member of every unit in the input is checked and the appropriate Phonetic Error Detection and Correction procedures are executed.

If a member of a unit does not have a corresponding procedure to execute, i.e., the symbol cannot have any phonetic error then it is explicitly validated. A symbol may also be explicitly validated by a detection and correction procedure.

The various Phonetic Error Detection and Correction procedures also attempt to detect possible inflected words. Whenever a root word is attached to a valid inflection, certain changes occur in the **boundary unit** of the root and the inflection. An analysis of the possible verbal inflections has shown that the presence of certain symbols in a word indicates the possibility that the given input may be an inflected word.

A procedure is called for execution when a particular symbol or one of the symbols in the associated Phonetic Equivalence Group appears in the input word. These procedures search appropriate dictionaries

for detection and correction of phonetic errors. After a successful search operation the input word may be detected as a Homophone. A procedure is called to retrieve, display and print the Homophone words and their meanings.

If the search operation is not successful, the procedure may attempt a modification on the input word on the basis of certain spelling rules. The modified word is searched in the same dictionary under a different Index Record or may be in another new dictionary. Finally, if the modification is not successful, either it will be kept or the original symbol may be restored. It depends on the spelling rules as well as on the nature of the words in Bengali. Sometimes, corrections are effected by a Phonetic Error Detection and Correction procedure.

8. Conclusion

The present report is basically on the detection and correction of isolated phonetic and homophone errors in Bengali. Currently, investigation is going on to detect and correct general spelling errors in Bengali using the same dictionary design. Development of context-based correction techniques in Bengali is a major challenge to be taken up in the near future.

9. REFERENCES

[1] Sinha R. M. K., Singh K. S.; " A Programme for Correction of Single Spelling Errors in Hindi Words"; JIETE, Special Issue on Computer Applications of Indian Languages & Scripts, Vol. 30, Number 6, pp. 249-251, 1984.

[2] Grasz, Jones, Webber (Ed.) ; Readings in Natural Language Processing ; Morgan Kaufman Publishers Inc., pp. 539 – 544, 1986.

[3] Srinivas B. , Ravisankar P. V.; " Recognizing Ill - formed Words in Indian Languages " ; Frontiers in Knowledge - Based Computing ; Bhatkar, Rege (Ed.); Proceedings of KBCS - 90, India, pp. 319 – 328, 1990.

[4] IS 13194 : 1991 ; Indian Standard Script Code for Information Interchange - ISCII; Bureau of Indian Standards ; 1991.

[5] Ranade Shreekant M. & Bhaskararao P.; " Marathi Word Identification Procedures : Applications for a Spelling Checker "; Proceedings of CPAL - 2; India, pp. 80 – 86, 1992.

[6] Bhatt Shashank , Doctor Raymond, Bhaskararao P.; " Implementation of a Spelling Checker for Hindi : Usage of Word Formation Rules and Economy " ; Proceedings of CPAL - 2; India, pp. 87 – 92, 1992.

[7] Kukich K., " Techniques for Automatically Correcting Words in Text ", ACM Computing Surveys, Vol. 24, No. 4, pp. 377 – 439, 1992.

[8] Chaudhuri B. B. , Pal U. ; " OCR Error Detection and Correction of an Inflectional Indian Language Script " ; 13th International Conference on Pattern Recognition, Vienna, 1996.

[9] Chaudhuri B. B. and Pal T., " Detection of Word Error Position and Correction Using Reversed Word Dictionary '', Proceedings of the International Conference on Computational Linguistics, Speech and Document Processing (ICCLSDP - '98), India, pp. C-41 - C-46, 1998.

[10] Bandyopadhyay S., " Studies on Detection and Correction of Spelling Errors in Bengali with a New Orthographic Dictionary using a Modified Script Processing Technique '', Ph. D. (Engg.) Thesis, Jadavpur University, 1998.

[11] Bandyopadhyay S., " A New Design for Orthographic Dictionaries in Alphabetic Languages '', Accepted for the 13th Pacific Asia Conference on Language, Information and Computation (PACLIC 13), Taiwan, R.O.C., February 10-11, 1999.

[12] Pabitra Sarkar, " Baanglaa Baanaan Sanskaar : Samasyaa O Sambhabanaa ", [a book in Bengali], Chirayaat Prakaashan Private Limited, Calcutta, India, 1987.

[13] Maahbubul Haq, " Baanglaa Baanaaner Niyam ", [a book in Bengali], Jaatiyo Saahitya Prakashani , Dhaka, Bangladesh, 1995.