Design and Implementation of a Morphology-based Spellchecker for Marathi, an Indian Language

Veena Dixit, Satish Dethe, Rushikesh K. Joshi
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai-400076, India
{veena, satishd, rkj}@cse.iitb.ac.in

Abstract

Morphological analysis is a core component of Technology for Indian languages. Complexities involved in spellchecking of documents in Marathi, an Indian language are described. Issues for both orthography and morphology are discussed. We have applied morphological analysis to a large number of words of different parts of speech. A spellchecker based on this analysis has been developed. The architecture of the spellchecker and the spell-checking algorithm based on morphological rules are outlined.

1. Introduction

Words can be defined from various perspectives such as phonological, morphological, grammatical, lexical, semantic, syntactic, orthographic, sociological and psycholinguistic (Dixon, 2004). The spellchecker's input is text, i.e. a stream of orthographic words. The perspectives used for spellcheckers and grammar checkers differ. The former are primarily based on vocabulary, while the latter require grammar rules. Spellcheckers may also use rules to reduce the size of vocabulary. A rule-based approach for spellcheckers is preferred for pan-Indian languages due to their morphological richness (WILSD, 2002). For Indian languages such as Marathi and Hindi, dictionaries covering all possible inflections, derivations and compounds obtainable from all root words do not exist. Not all Marathi words in frequent use are stored in the dictionary. For example, for a single noun in Marathi, over 200 forms that are either adjectives or adverbs may be possible. Similarly, a verb may exhibit over 450 forms. At the same time, the language is expected to include over 10,000 nouns and over 1,900 verbs. Over 175 postpositions can be attached to nominal and verbal entities. Some postpositions can occur in compound forms with most other postpositions. In addition, there are many kinds of derivable words such as causative verbs like *karavane,* i.e. 'to make (someone) to do (something)', which is derivable from root *karane* i.e. 'to do', and abstract nouns like *gharpan* i.e. 'homeliness', which is derivable from *ghar* i.e. 'home'. Marathi has tendency to use onomatopoeic words frequently, which are not maintained in the dictionary. The rich morphological nature of the language makes a morphology-based approach more suitable. Also as Marathi corpora in electronic media is not available so far, possibility of a corpora-based spell-checker was ruled out. A morphology based spellchecker has other advantages such as its ability to handle the *name-identity* problem, i.e. it can absorb new words and foreign words that are not included in the dictionary. New words may be absorbed by categorizing them into appropriate paradigms. Further, the approach can be drawn upon in building grammar checkers. A morphological rule base developed for spellchecker is also a stepping-stone for natural language processing.

We discuss the architecture and implementation of a rule-based spellchecker for Marathi, a major Indian Language. To our knowledge, this is the first major initiative for morphology-based spellchecking for Marathi. The spellchecker is based on the rules of morphology (Damale, 1970; Pandharipande, 2000) and the rules of orthography (Govt. of Maharashtra, 1986; Gokhale, 1993; Phadke, 2001). Morphological rules address word categories and their possible inflections.

The next section discusses issues related to rules of orthography. Morphological issues for various word categories are discussed in Section 3. An implementation and its evaluation are provided respectively in Sections 4 and 5. In most places, IPA is used to represent characters in Marathi.

2. Some Orthographical Issues

Marathi is written in Devanagari script. It maps the phonemic shape (phonemes and their sequence) of a word to Devanagari symbols through more or less one to one mapping. A spellchecker for Marathi has to consider the symbols for 34 *vyanjans* (consonants), 15 *swaras* (13 vowels, nasalization and aspiration) and 15 *matras* (vowels, nasalization, aspiration and *halant* markers) (Damale, 1970). Twelve *matras* are used to indicate the presence of a particular vowel at respective position in the phonemic representation of the word. A special *matra* called *halant* represents absence of phoneme 'schwa' instead of indicating presence of it. Schwa is latent in consonantal alphabet. Besides these symbols, over 180 *cluster characters*, commonly occurring mathematical symbols and punctuation marks are considered.

An *alphabet* represents a phonemic sequence <consonant, 'schwa'> as noted in (Wakankar, 1968). A cluster character may be formed by one of the two sequences <consonant, alphabet> and <consonant, consonant, alphabet>. Following combinations occur as *characters* in a written script: an independent vowel, an independent consonant, an independent cluster character, sequence <alphabet, matra> and sequence <cluster character, matra except *halant*>. Valid combinations are defined by the rules of orthography, which in turn are based on etymology (Gokhale, 1993) and phonemic sequences of words (Damale, 1970). A spellchecker that considers these factors can automatically reject certain invalid sequences and suggest alternatives or autocorrect some of them (Joshi, 2002).

The rules of morphology need to capture changes in phonemes. These are represented as transformations of *matras* representing corresponding vowels. However, when vowel *schwa* combines with a consonant, no separable matra appears in the corresponding alphabet in most encodings used today due to latency of schwa in Devanagari. With such encodings, transformations of type (schwa→matra) or (matra→schwa) cannot be handled directly at encoding level. For example, in morphological transformation of word राम (ramə) to word रामाला (ramala) the rule (schwa →ा) is applied on alphabet म (m). However, in Unicode representation of the word राम (ramə), vowel schwa is absent. Similarly, rule (matra ू→schwa i.e. अ (ə)) is applied on alphabet ड in transformation of word लाडू (laɖu) to word लाडवाला (laɖəvala), while *schwa* does not occur in the Unicode representation of the word. The spellchecker needs to analyze the word from orthographic point of view by applying the orthographic rules given above. Interestingly, this problem does not arise in IITK mapping for Devanagari, which uses English alphabet for transcription. The mapping uses character 'a' to capture vowel *schwa*. Hence, IITK mapping was chosen to implement morphological rules in the spellchecker.

If the ultimate vowel in a word is *schwa*, the penultimate vowel is usually written in its long form. In such cases, after morphological transformations, long penultimate vowel (ू or ी, i.e. U or I) in the root word is transformed to short vowel (ु or ि, i.e. u or i) if the vowel is retained in the transformation. Govt. of Maharashtra (1986) has standardized various rules of orthography for contemporary Marathi.

3. Rules of Morphology

Morphological analysis is applied to the categories of nouns, pronouns, adjectives, verbs, adverbs, postpositions, conjunctions and interjections. In Marathi, it is convenient to use rules of replacement to capture all types of morphological behavior including those captured in examples given below.

- Changes to a word's phonemic shape at the end of the word considering the latent schwa as in transformation of राम (ramə) to रामाला (ramala) as discussed above.
- Changes to a word's phonemic shape not only at the end of the word but anywhere in the middle of the word as in transformation of खातापिता (k^hatapita) to खात्यापित्या (k^hatyapitya).
- Changes to all vowels in the phonemic shape of the word such as in transformations of ऊ (u:) and मूल (mu:lə) to उवे (uve) and मुला (mula) respectively.
- Other examples include deletion of ultimate or penultimate consonant, addition of a consonant and vowel pair at the end of the word.

Rules of replacement are generic enough to also cover all possibilities of additions and deletions of consonants and vowels. Replacement rules consider latent schwa and null components as and when required.

In Marathi, postpositions are attached to oblique forms of nominal and verbal entities. Hence, postposition morphology is important for morphological analysis of these categories. Most of the rules can be expressed in the form of transformation tables. Order of suffixes is captured through additional syntactic rules. Over 13,000 root words have been collected and classified by part of speech. For each word category, analysis was performed to derive inflectional morphological rules. Primarily, the parameters that were considered are tense, aspect, mood (TAM) and gender, number, person (GNP) and attachment of postpositions.

3.1 Postposition Morphology

Paradigms of postpositions are created based on their linguistic behavior. They include case markers (vibhakti pratyay) and a class of postpositions called *shabdayogi avyay.* The latter are attached to singular and plural forms of nouns and pronouns. Some shabdayogi avyays exhibit specific behavior. For example, some postpositions need to be written separately when they follow syllable च्या (cya), which is a case marker. Some shabdayogi avyays can be suffixed with case markers चा (ca), ची (cI), चे (ce), च्या (cya). Some shabdayogi avyays can be composed of others. Postpositions ही (hI) and च (cə) can be attached before some shabdayogi avyays, but not before vibhakti pratyays. Some shabdayogi avyays can be attached to different oblique forms of verbs. Currently, the spellchecker handles the first level of postpositions in the above classification.

3.2 Noun Morphology

Changes due to the attachment of postpositions are different for singular and plural forms of nouns. The changed form of a noun to which such attachment is done, is called Saamaanyaroop (oblique form) of that noun. For example, in morphological transformation of word राम (ramə) to word रामाला (ramala), the samanyaroop of राम (ramə) is रामा (rama). Table 1 represents a snapshot of possible paradigms of inflections in nouns.

3.3 Pronoun Morphology

Exhaustive list of all possible (over 550) inflections of all pronouns is prepared because pronouns show very irregular behavior. The ratio of inflectional rules to

Changing part				Change											
				Feminine											
pc	pv	uc	uv	sso				spf				spo			
				pc	pv	uc	uv	pc	pv	uc	uv	pc	pv	uc	uv
प्प Pp	अ ə	ल l	अ ə	प P	अ ə	ल l	ए e	प P	अ ə	ल l	आ a	प P	अ ə	ल l	आं Ã
-	-	स s	अ ə	-	-	श ʃ	ई I	-	-	श ʃ	ई I	-	-	श ʃ	ई Ĩ

sso: suffix for singular oblique form *pf: suffix for plural form* *spo: suffix for plural oblique form*
pc: Penultimate consonant *pv: Penultimate vowel* *uc: Ultimate consonant*
uv: Ultimate vowel.

Table 1: Snapshot of Noun Morphology

actual forms in the case of pronouns is close to one. A pronoun has a specific single oblique form to which all s*habdayogi avyays* are attached.

3.4 Verb Morphology

Aakhyaata Theory is the basis of verb morphology analysis. It systematically segments the verb forms into verb roots and terminating suffixes called *Aakhyaatas*. *Aakhyaata* represents information about TAM and GNP. They are named according to the phonemic shape such as *taakhyaata, vaakhyaat* and *laakhyaata*. A regular verb root generates over 80 forms. In addition to regular verbs, there are over 35 irregular verbs. The rules are represented in the form of tables.

3.5 Adjective Morphology

Adjectives are classified in inflectional and non-inflectional categories. Inflections result from gender, number and attachment of postpositions to the noun modified by such adjective. Table 2 shows a snapshot of inflectional rules. In the spellchecker, the root form is chosen as masculine form, from which other forms are generated.

Changing part in masculine form	Change		
	Feminine	Neuter	Oblique form
आ a	ई I	ए e	या ya

Table 2. Adjective Morphology

When genitive case markers or some *Shabdayogi avyays* are attached to nouns, it produces adjectives. These forms are automatically covered in noun morphology.

3.6 Adverb, Conjunction and Interjections

This is an important class of part of speech, for which the rule-based approach proved to be appropriate. Attachment of postpositions to nouns, verbs and pronouns is one of the strategies of adverb formation. In addition, there are non-inflectional adverbs. The set of derived adverbs is automatically covered at the level of morphology of postpositions, nouns, verbs and pronouns. The list of all lexicalized adverbs is constructed. Similarly, all conjunctions and interjections are handled as a list since they are non-inflectional. When some postpositions are attached to demonstrative pronouns, conjunctions are derived. These are handled at the level of rules for pronouns and postpositions.

4. Implementation

Figure 1 illustrates the architecture of the spellchecker. Using the services offered by spellchecker's interface (SCI), the front end of the system provides spellchecking facilities for Marathi documents in IITK, UTF-8 and Phonetic formats. A font converter is supported to process convert documents in other formats to IITK format which is used in the spellchecking process. Unicode is used for the display unit. The front end provides support for text editing, storage format conversion, highlighting of invalid words and handling of user actions on them. A highlighted word can be ignored, replaced or can be added to user's vocabulary. Alternatives are suggested based on a string distance (Soukoreff, 2001) and morphological rules.

The SCI consults the Morphology Analyzer (MA), which in turn consults individual part of speech analyzers for noun, adjectives, verb and other categories. The individual part of speech analyzers use their independent rule bases as shown in the figure. Besides, a user level wordlist can also be plugged in.

The algorithm to check the validity of a word is outlined below.

1) If the word w is not found as it is in the vocabulary, proceed to step 2, else accept the word and terminate.

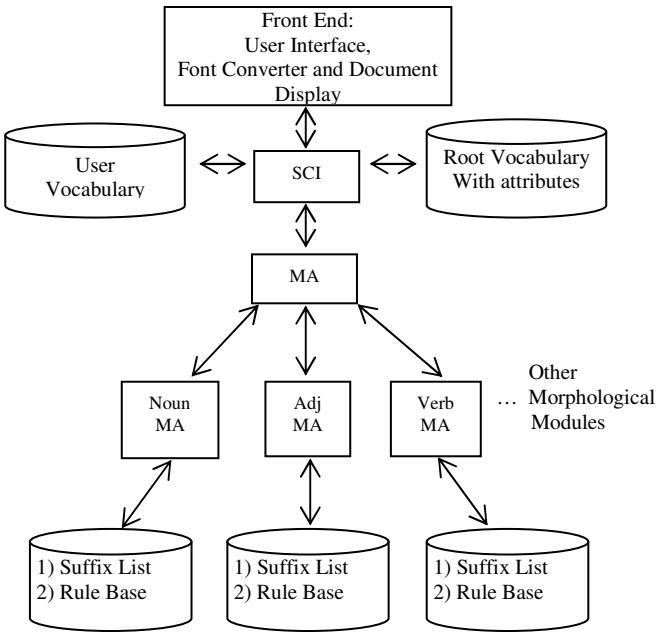

Figure 1: Architecture of the Spellchecker

2) Scan the word w from right to left to identify a valid suffix string 's_2' such that s_2 occurs in at least one rule of the form ($s_1 \rightarrow s_2$). Note that s_1 and s_2 may be of length more than 1, and s_1 may be a substring in s_2. If such a rule is not found, reject the word as invalid and terminate, else proceed to step 3.

3) At the rear end of the word, carry a transformation ($s_2 \rightarrow s_1$) to obtain pruned word w_1 from w. If the transformed word w_1 is found in vocabulary and if the rule ($s_1 \rightarrow s_2$) is applicable for the word class of w_1, accept w as valid word and terminate; else proceed to step 4.

4) Go to step 2 to find another applicable rule.

If the word is detected as invalid, suggestions are provided based on left to right matching supported by inflectional rules and a string distance (Soukoreff, 2001). Besides morphological analysis, the spell-checker also considers the rules of orthography as discussed in Section 2. The Spellchecker is implemented in Java. It supports documents in IITK format. For display, the documents are converted into Unicode. The rules and the word lists with attributes are also stored in IITK format.

As an example, consider input word *rAmAkaravI* in IITK format. Since the word is an inflected word, the algorithm proceeds to step 2. In step 2, a rule (a → A karavI) is found. Applying step 3, pruned word rAma is obtained. Since this word is found in the vocabulary, the applicability of this rule to the class of the word is examined. In this case, the rule being applicable, the input word is declared valid.

5. Evaluation

A manual analysis of 10,648 words from a corpus, which were declared by the spellchecker as valid showed that 46 words among them were invalid. This implied an accuracy of validity of 99.57%. The reasons of error were traced to missing implementation of rules and exceptional cases. Similarly, a manual analysis of words declared as invalid showed that a large percentage of words were wrongly identified as invalid. The reasons were traced mainly to incomplete vocabularies and also to multiple ordered suffixes which have not been handled in the current version. The

current size of the vocabulary is limited to about 13,000 words. Enhancement in the vocabulary will improve the accuracy.

Various kinds of errors that can occur include misspelled root word and misspelled or inappropriate suffix and wrong order of attachment of multiple suffixes. Suggestions for words found to be incorrect are provided by considering the word's three constituents, which are root, stem forming suffix and case marker or postposition. A right to left (depth first) strategy is used to locate all possible correct formulations. A suggested formulation is allowed to differ at most by one vowel and one consonant. Finally, all suggestions are sorted based on string distance and first eight suggestions are displayed. It was found that in most of the cases that were tested this scheme resulted in obtaining the expected word in first three suggestions if the input word is misspelled by a vowel and/or a consonant.

6. Conclusion and Future Work

Morphological analysis on over 40,000 Marathi word forms was performed for different part of speech categories. As typical to Indian Languages, the possible inflections of a single word are huge in number. Some challenges in building a spellchecker for handling such complex linguistic phenomenon were discussed. A spellchecker architecture and implementation for first level suffixes based on morphological analysis and rules of orthography was presented. Initial tests showed that the approach was very accurate in declaring words as valid. Further enhancements of derivational morphology will help in increasing the vocabulary. Besides enhancing word lists and rules, enhancements for representing rules for ordering of multiple suffixes in all part of speech categories are required. More elaborate orthographic rules need to be incorporated. Morphology based spellchecker may be extended to include further syntactic and semantic analysis. Besides spellchecking, the morphology based analysis is currently being used in a few applications at the Center for Indian Languages. The morphological analysis of a word serves as a foundation for POS-tagging. Similarly, it is being used in stemming for searching root words in Marathi Wordnet, as well as in AQUA, a Marathi search engine.

7. References

Damale M.K., 1970. *Shastriya Marathi Vyaakarana*, Deshmukh and Company, Pune.

R.M.W. Dixon, 2003. *Word: A Cross – Linguistic Typology*, Press Syndicate of Cambridge, Cambridge, U. K. 2004.

Pandharipande R., 2000. *Marathi*, Routledge Publication. *Marathi Shuddhalekhanaache Niyam*, 1986. A Publication of Govt. of Maharashtra.

Gokhale D.N., 1993. *Shuddhalekhan Vivek*, Soham Prakashan, Pune.

Phadke Arun, 2001. *Marathi Lekhan Kosh*, Keshav Bhikaji Dhavale Publishers, Mumbai, 2001.

Wakankar L. S, 1968. *Ganeshvidya*, Script Study Group. *Proceedings of Workshop on Indian Language Spell-checker Design (WILSD),* 2002. Resource center for Indian Language Technology Solutions, Indian Statistical Institute, Calcutta, July 18-19, 2002.

Soukoreff R.W. and MacKenzie I.S., 2001. Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic, *CHI'01 Extended Abstracts on Conference on Human Factors in Computing System*, pp. 319-320.

Joshi Rushikesh K. and Hemant Patil, 2002. Low Level Syntax Checker and Auto Corrector, presented at TDIL meet, Ministry of Information Technology, Delhi.

Acknowledgements: First two authors were supported through a grant from Ministry of Information Technology under TDIL project. The authors are thankful to Pushpak Bhattacharyya and members of CFILT for valuable comments.