Processing of Kridanta (Participle) in Marathi

Ganesh Bhosale, Subodh Kembhavi, Archana Amberkar, Supriya Mhatre, Lata Popale, Pushpak Bhattacharyya
Department of Computer Science and Engineering,
IIT Bombay

pb@cse.iitb.ac.in

Abstract

In this paper, we propose and evaluate a Finite State Machine (FSM) approach for *Krudanta* processing for Marathi; an agglutinative and highly inflectional language belonging to the Indo-European family with semblances of Dravidian languages. *Krudanta* is the phenomenon of participial construction that needs adroit handling for high quality output in machine translation (MT). We contextualize this work in a transfer based Marathi to Hindi Machine Translation system to observe the importance of *Krudanta* processing. Results with and without *Krudanta* processing establish the need for this work.

1 Introduction

Marathi morphology makes use of agglutinative, inflectional, and analytic forms. A specific feature of the syntax is the widespread use of participial constructions to express subordinating relations as in Dravidian languages. The speaker population of Marathi all over the world is close to 70 million.

Krudanta is a term used in traditional Marathi grammar for those derivational morphemes, which are affixed to verbal roots in order to derive nouns, adjectives and adverbs; as opposed to *Taddhita* suffixes which are affixed to nouns and which derive words belonging to the remaining 3 grammatical categories. Any word affixed with a *Krudanta* suffix is called ***Krudanta***.

Krudanta are complex forms to process. These words frequently occur in the Marathi language (about 15%). Marathi *Krudanta* are used in place of some particular types of relative clauses. In the sentence 'the boy who swims everyday' is expressed as 'रोज पोहणारा मुलगा' {roj pohanaaraa mulagaa}, पोहणारा {pohanaaraa} {one who swims} is a *krudanta* which is used frequently rather than the less frequent expression 'रोज पोहतो तो मुलगा' {roj pohatoo to mulagaa} {daily swims that boy}. In the context of Marathi to Hindi Machine Translation, we experienced a number of challenges in processing of *Krudanta* forms and observed its importance in getting higher level of performance in Machine Translation. *Krudanta* are meaning bearing units in the sentence. Hence, processing of *Krudanta* forms is an important task.

1.1. Related Work

Morphological Analyzers for various languages have been studied and developed for years. Eryiğit and Adalı (2004) propose a suffix stripping approach for Turkish. Many Morphological Analyzers have been developed using the two-level morphological model (Oflazer, 1993; Kim et al., 1994, Antworth, 1991). This model proves to be very useful for developing the morphological analyzers for agglutinative languages. Dixit et al. (2006) developed a morphological analyzer with the purpose of using it for spell checking. Though their analyzer successfully analyzes the words with a single suffix, its scope is restricted to the han-

Proceedings of ICON-2011: 9th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2011

dling of only first level suffixes for simple word forms.

The work on *krudanta* processing takes place in the context of a consortium project on Indian Language to Indian Language Machine Translation (ILMT)[1] system for nine Indian language pairs. The system is based on the Analysis-Transfer-Generation paradigm.

The roadmap of the paper is as follows. In section 2, we discuss the *Krudanta* Theory for Marathi and a few participial construction examples from other agglutinative languages. In section 3, we describe the computational challenges of processing *Krudanta* forms. Section 4 is on the Morphological Processing of *Krudanta* forms. In section 5, we describe the role of Chunker and Lexical Transfer module in *Krudanta* processing. In section 6, we evaluate the system with a discussion on error analysis. Finally in section 7, we conclude the paper with pointers to future work.

2 *Krudanta* Theory

Krudanta forms are a type of forms obtained by affixing derivational morphemes to the base morpheme. One of the main and defining characteristics of derivational morphemes is that they change the grammatical category (part of speech) of the base morpheme they are affixed to. For example the Marathi word तरुणपणा, 'tarunpanaa' meaning 'the state or quality of being young', is a noun made of two morphemes 'tarun', which is an adjective meaning 'young', and 'panaa', which is a derivational affix used to create abstract nouns.

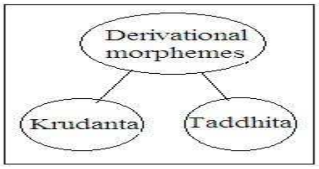

Figure 1: Derivational Morphemes

Krudanta forms are the forms derived from verbs. *Krudanta*s can be nouns, adjectives and adverbs. Examples of these are given in the following section. Figure 1 shows the other derivational form that is common in Marathi, viz., the Taddhita form which are derived from nouns.

Words derived from verbal roots can be classified as per their functions as nouns, adjectives and adverbs. Some examples of each of these classes are as follows:

- **Nouns derived from verbs:**

1. वाच {vaach} {read} derives वाचणे {vaa-chaNe} {in the act of reading}.
2. उतर {utara} {climb down} derives उतरण {utaraN} {downward slope}.

- **Adjectives derived from verbs:**

1. चाव {chav} {bite} derives चावणारा {chaa-vaNaara} {one who bites}.
2. खा {khaa} {eat} derives खाल्लेले {khallele} {something that is eaten}.

- **Adverbs derived from verbs**:

1. पळ {paL}{run} पळताना {paLataanaa} {while running}
2. बस {bas}{sit} बसून {basun}{manner adverb of sit}

Dixit et al (2006) describe 8 types of Krudantas which are classified on the basis of the particular derivational morpheme affixed to the base. Various categories of derived words are listed in Table 1.

Some of the Dravidian languages have similar participial suffixes as in Marathi:

Kannada: In the sentence:
muridiruwaa kombe jennu esee
Broken to branch throw
Throw away the broken branch.

Here, muridiruwaa is a participial form (here, an adjective derived from a verb), which is similar to the Marathi *Krudanta* forms derived by affixing the derivational morpheme 'lela' to the verbal base.

Telugu: In the sentence:
ame padutunnappudoo nenoo panichesanoo
she singing i work
I did work while she was singing.

[1] http://sampark.iiit.ac.in

***Krudanta* Type**	**Example**	**Aspect**
Ne	वाचण्यासाठी {vaachaNyasaThI} {for reading}	Perfective
La	वाचल्यावर {vaachalyavar} {after reading}	Perfective
Tana	वाचताना {vaachataanaa} {while reading}	Durative
Lela	वाचलेले पुस्तक {vaachalele pustak} {book which is/was read}	Perfective
Nara	वाचणारा {vaachaNaaraa} {the one who reads}	Stative
Va	वाचावे {vaachave} {must read}	Inceptive

Oon	वाचून {vaachun}
	{having read}
		Completive

Table 1: *Krudanta* Types in Marathi

2.3 Kridanta forms in other Agglutinative Languages

padutunnappudoo is a participial form (here, an adverb derived from a verb), which is similar to the Marathi *Krudanta* forms derived by affixing the derivational morpheme 'taanaa' to the verbal base.

Turkish:
hazirlanmis plan
prepare-past plan
The plan which has been prepared
Here, *hazirlanmis* is the participial form derived from a verb.

3 Challenges in Marathi Morphological Processing

Marathi is a morphologically rich language with a high level of agglutination. Inflections can get attached at various positions. We mention some of these challenges.

3.1 High level of inflection and agglutinative morphemes

There are many examples of *Krudanta* forms with more than two suffixes attached to the root after inflection. Stacking of affixes is common. Consider in this example, मारणाऱ्यानीदेखील {maarNaaryaneedeKeel} {the killers also} {मार + णारा + नी + देखील} the root is मार {maar} {kill} as a verb attached with three suffixes णाऱ्या, नी and देखील respectively. Here both the suffixes णाऱ्या and नी are followed by another morpheme. Morphological analysis with affix stripping (Eryiğit and Adalı, 2004) fails to process such a *Krudanta* form.

3.2 Complex Forms of *Krudanta*

As *Krudanta* are made up of verbal roots, they sometimes encode 'aspect' feature. Hence it is very difficult for the traditional affix stripping approach without a proper rule to process such complex forms. But, maintaining many such rules for each possible grammatical form is not feasible. For example in केलेल्यानेसुद्धा {kelelyanesudhaa} {the person who did as well}, कर is the root and is affixed with the *Krudanta* morpheme 'lela', which denotes ‘perfective’ aspect. The emergent form is then affixed with two more suffixes ने and सुद्धा respectively.

3.3 Ambiguity in Finite and Non-Finite forms

In Marathi, there are pairs of morphemes that have similar phonological and orthographical shape and that bring about similar changes in the phonological and orthographical shape of the base word they are affixed to. As a result, the final forms are phonologically and orthographically similar, but have two different meanings. For example, there can be two Marathi morphemes, which are represented by the letter ‘त’ {t}, one of them being an inflectional morpheme denoting habitual past and the other a derivational morpheme denoting progressive aspect. Thus, when attached to a verbal root like ‘फिर’ {fir} meaning ‘to wander’, these two suffixes produce two similar forms- फिरत {phirat}, one of which means ‘they used to wander’, while the other means ‘wandering’. In such cases, the Morphological Analyzer should be able to produce both the analyses. After that, we need to disambiguate these analyses, by using statistical HMM along with distinguished morph features (Hakkani, 2002).

3.4 Two consecutive *Krudantas*

Consider, for example, the individual *Krudanta*s चालून {chalun} and येताना {yetaanaa}, which mean ‘after walking’ and ‘while coming’ respectively. However, when they are used together चालून येताना ‘chalun yetaanaa’, the meaning is {while coming walking}. A sequence of *Krudanta*s may have a meaning different from the constituent *Krudanta*s. In case of two such consecutive *Krudanta*s, we get the chunk type as VGNF (Verb Group Non-Finite). In the above example, the individual processing of each *Krudanta* gives an accurate translation, but for two consecutive *Krudantas* the translation process fails because of the incorrect Vibhakti (suffix generation after head computation in the chunk).

4 Morphological Processing of Marathi keeping *krudanta* in focus

We have developed a Morphological Analyzer for Marathi using a finite state machine approach. For that purpose, we are using SFST (Stuttgart Finite State Transducer) tool [2]which is used for the implementation of computational morphology. This approach builds on the work of Bapat et al. (2010). There are three important components in the architecture of the morphological analyzer as shown in Figure 2.

Inflector: The inflector takes as input lexicon and suffix replacement rules (SRR). The format of an SRR is (*Paradigm, Suffix, Ultimate Insertion, Ultimate Deletion, Penultimate Insertion, Penultimate Deletion, and morph features*). Consider, for example, मुले {mule} {boys} where to extract the root मूल we used the SRR rule below <मूल>< , ◌े, ◌ू, ◌ु><noun,n,pl,,,,,d>. This means the following: as a first step, we remove ◌े from ultimate position of the word and then we replace ◌ू for ◌ु in the penultimate position to get the root word मूल. The Inflector creates a flat file where it stores all the inflected forms of words present in the lexicon along with the associated grammatical attributes and labels. We call this flat file as a dictionary of inflected forms.

Morphological Recognizer: The Morphological Recognizer breaks a given word into morpheme segments. Morphotactics are modeled using the FSM which carried out the segmentation with the help of generated ‘dictionary of inflected forms’.

Morphological Parser: These segments along with their labels are input to the Morphological Parser which uses the ‘dictionary of inflected forms’ to extract the root and also gives the morphosyntactic properties of each morpheme.

[2] http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

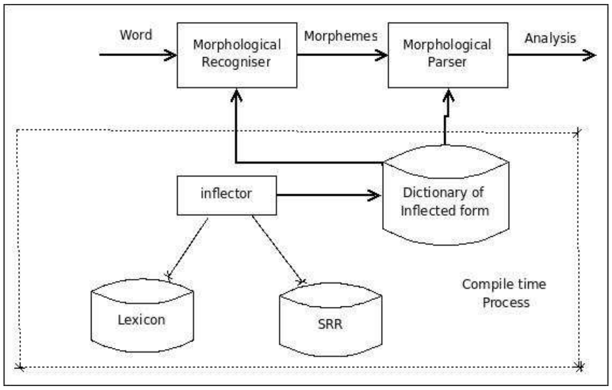

Figure 2: Architecture of the Marathi Morphological Analyzer

We show an example of *Krudanta* in figure 3. Notice how the large number of morphemes get isolated through the above mentioned stages.

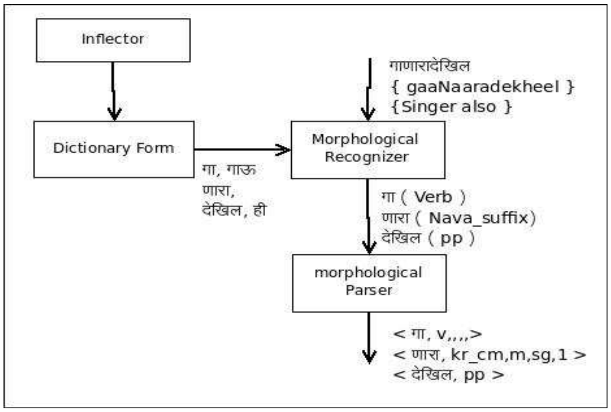

Figure 3: *Krudanta* Processing Example

4.1 FSM Rules for *Krudanta*

Figure 4 shows the grammatical rules for *Krudanta* processing in the form of a finite state machine. We have VERBS, VERBS_LE, *etc.* files which contains possible verbs list. FSM rules are constructed depending on the all possible *Krudanta* forms. These rules are used by the Morphological Recognizer to break a given word into morphemes.

Consider the word गाणारादेखिल {gaanaaradekhil}{singer also} which is the *Krudanta* where गा is the root and 'णारा' and 'देखिल' are the suffixes attached.

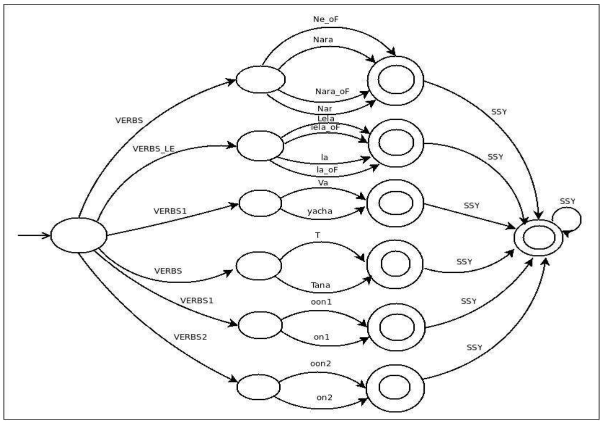

Figure 4: FSM rules for *Krudanta*

To process this *Krudanta* word the following FSM rule is used. [VERBS -> Nara_DF -> SSY]. In processing of गाणारेदेखिल {gaanaaredekhil} {singers also} (which is the plural form of the गाणारादेखिल) following FSM rule is used. [VERBS -> Nara_OF -> SSY] where VERBS file contains verb roots like गा, Nara_DF file contains suffix णारा, Nara_OF file contains suffix णारे with other inflected forms of suffix 'णारा', and SSY file contains suffixes like देखिल.

5 Marathi Hindi MT and *Krudanta* Processing

The Marathi to Hindi MT system processes one chunk at a time instead of processing each token. Compound and conjunct verbs are grouped as a single chunk and tagged as Verb group finite (VGF), e.g., लिहून टाकला {lihun Taakalaa} {Finished writing}. Such chunks contain one or more tokens which are marked as *Krudanta*s by the morph analyzer. e.g. 'लिहून 'in the above example would get marked as an '*oon_Krudanta*' by the morph analyzer. However, since the chunk is a finite verb group, the *Krudanta* type of such tokens needs to be ignored and the chunk is to be transferred like finite verb forms. Thus the Chunker plays a crucial role in *Krudanta* transfer. It tags finite verb groups as VGF chunk even if they contain a token which could be detected as a *Krudanta*. Without such chunking, a token which is part of a finite verb group could get processed as *Krudanta* and this could lead to an incorrect translation.

6 Experimental setup to evaluate *krudanta* processing

We have tested seven types of *Krudantas* on 1000 possible *Krudanta* forms. We have performed two types of evaluation. In the direct evaluation, we considered the morphological features generated after *Krudanta* processing using MA and in the indirect evaluation, we analyzed the Marathi-Hindi machine translation outputs when such *Krudanta* forms were present in the sentence.

In the indirect evaluation, we have evaluated the Marathi-Hindi MT system to illustrate the importance of *Krudanta* processing. In the first set of experiments we got the accuracy figures for MT, using the affix stripping based MA with no special attention to *krudanta*. Next we got these figures with the finite state based MA with focus on *krudanta*. These results were then compared.

6.1 Direct Evaluation:

Direct evaluation is based on the morphological analysis of a given *Krudanta* form. We evaluated the system by calculating precision and recall for *krudanta* words.

C = Number of Correct analysis, M = Number of Missed analysis, W = Number of Wrong analysis

Precision= C/(C+W)

Recall= C/(C+M)

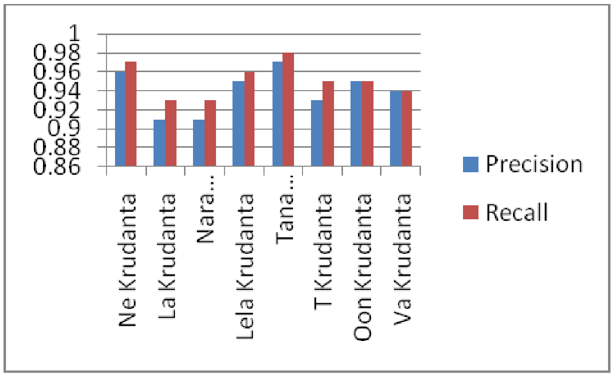

Figure 5: *krudanta* processing: Precision and Recall

As figure 5 shows, the *krudanta* forms are analysed with an average accuracy of 95%.

6.2 Indirect Evaluation:

To evaluate the system performance for *Krudantas*, we used the Marathi to Hindi Machine Translation system based on transfer approach. We gave 1000 different web sentences to Marathi to Hindi Machine Translation system which includes all *Krudanta* forms. The accuracy is measured on the score given by linguists, based on the Hindi translation quality (table 2).

Score : 5	Correct Translation
Score : 4	Understandable with minor errors
Score : 3	Understandable with major errors
Score : 2	Not Understandable
Score : 1	Non sense translation

Table 2: Score based on Hindi MT quality

S5: Number of score 5 sentences, S4: Number of score 4 sentences, S3: Number of score 3 sentences N: Total Number of sentences

Accuracy= (1*S5+0.8*S4+0.6*S3)/N

We calculated the accuracy for the sentences after integrating MA (developed using affix stripping approach) into the MT system. Again on the same set of the sentences, we calculated the accuracy after integrating MA (developed using FSM approach) into the MT system. The results are shown in table 3.

Approach	MT accuracy
Affix stripping (no special attention to *krudanta*)	56.27 %
Finite State Machine (with special attention to *krudanta*)	63.45 %

Table 3: Machine Translation accuracy with and without *krudanta* processing

6.3 Results and discussions

Marathi-Hindi MT improved with *krudanta* processing. We observed that *Krudanta* forms with more than two suffixes and inflections are processed accurately by FSM based Morphological Analyzer. Some *Krudanta* forms of the type “Tana", "Nar" had high precision and recall because they have simpler forms to process, but some *Krudanta* forms of type "Nara", "La" had less than 85% precision because of their highly inflective nature. We found that in most of the complex Krudanta forms, the morphological analyzer was able to produce accurate feature analysis, but the final quality of translation depended on other modules like lexical transfer, word generator.

6.4 Error Analysis

Reason	Number of Errors
Grammar Rule Missing	4
Words not in Lexicon	6
Wrong Morph Features	3
Ambiguity in Krudanta type	4
Agglutinative Suffix	5
Two Token *Krudanta*	4
Word Generator	4

Table 4: Error Analysis for *krudanta* procssing

We found that 30 *Krudanta* words failed to be processed. Out of these, four failed because of missing grammar rules. Six failed because the word was not in lexicon. The words belonging to the "T" and "la" *Krudanta* category were ambiguous words which could be finite or non-finite verb forms like in example गेला {gelaa}{went}{Finite Form} and गेला {gelaa} {last one} {Non-Finite Form}. In the MT system, we observed that four consecutive *Krudantas* were not processed because of wrong Vibhakti generation like in the example, वाचून झाल्यावर {vaachun Zalyaavar}{after completion of reading}. We found *Krudanta* examples which had more than two suffixes attached were not processed in the lexical substitution module, though MA was able to process them accurately.

7 Conclusion and Future work

We presented a high accuracy morphological analyser for *Krudanta* forms using the Finite State Machine technique. The accuracy figures as high as 95% in direct evaluation and the performance improvement in Marathi to Hindi translation are observed. Our approach could be useful for Dravidian languages and agglutinative languages like Turkish and Hungarian. As future work, we would like to study how best the *krudanta* information can be passed onward to other MT modules like lexical transfer for high quality output generation.

References

Antworth 1990. *PC-KIMMO: A Two level Processor for Morphological Analysis.* Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, Texas.

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

Dilek Z. Hakkani-Tur, Kemal Oflazer and Gokhan Tur. 2002. *Statistical Morphological Disambiguation for Agglutinative anguages*. Computers and the Humanities, Kluwer Academic Publishers. Printed in the Netherlands.

Dixit, Veena, Satish Dethe,and Rushikesh K. Joshi. 2006. *Design and Implementation of a Morphology-based Spellchecker for Marathi, an Indian Language*. In Special issue on Human Language Technologies as a challenge for Computer Science and Linguistics. Part I. 15, pages 309–316. Archives of Control Sciences.

Eryiğit,Gülşen and Adalı Eşref. 2004. *An Affix stripping Morphological Analyzer for Turkish*. In IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, pages 299–304.

Bapat Mugdha, Harshada Gune and Pushpak Bhattacharya 2010. *A Paradigm-Based Finite State Morphological Analyzer for Marathi*. the 23rd International Conference on Computational Linguistics, Beijing, China, 2010.

Oflazer, Kemal 1993. Two-level Description of Turkish Morphology. In The European Chapter of the ACL (EACL).

Prószéky G., Kis B 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. ACL'99, Maryland.