Logistic Regression for Automatic Lexical Level Morphological Paradigm Selection for Konkani Nouns

Shilpa Desai
Department of Computer Science and Technology
Goa University
sndesai@gmail.com

Jyoti Pawar
Department of Computer Science and Technology
Goa University
jyotidpawar@gmail.com

Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT-Patna
pb@cse.iitp.ac.in

Abstract

Automatic selection of morphological paradigm for a noun lemma is necessary to automate the task of building morphological analyzer for nouns with minimal human interventions. Morphological paradigms can be of two types namely surface level morphological paradigms and lexical level morphological paradigms. In this paper we present a method to automatically select lexical level morphological paradigms for Konkani nouns. Using the proposed concept of paradigm differentiating measure to generate a training data set we found that logistic regression can be used to automatically select lexical level morphological paradigms with an F-Score of 0.957.

1 Introduction

Morphological analysis is required for many NLP applications such as Spell Checkers, Text to Speech Systems, Rule Based Machine Translation, etc. Finite State Transducers (FSTs) are ideal for developing Morphological Analyzer for a language because they are computationally efficient, inherently bidirectional and can also be used for word generation. FST based Morphological Analyzers are based on word and paradigm model, wherein a word lemma is mapped to a corresponding Morphological Paradigm. A Morphological Paradigm is used to generate all possible word forms for a given word lemma. To develop a FST based Morphological Analyzer two resources namely Morphological Paradigm List and Morphological Lexicon are required. Morphological Paradigm List is prepared referring to grammar books, morphology related linguistics thesis in Konkani and elaborate discussions with linguists. Lemmas[1] in the language are then mapped to appropriate Morphological Paradigms to create a Morphological Lexicon. Mapping of lemmas to Morphological Paradigms is time consuming when done manually.

Automating creation of Morphological Lexicon requires automatic mapping of lemmas to morphological paradigms. Morphological Paradigms can be defined at two level surface level and lexical level. At surface level two different morphological paradigms will generate a different Inflection Set for a given lemma whereas at lexical level two different morphological paradigms could generate same Inflection Set for a given lemma. Thus automatically choosing a correct morphological paradigm at the lexical level cannot be based on Suffix Evidence Value as in previous used methods (Carlos et al., 2009).

In this paper, we present the use of logistic regression for Automatic Lexical Level Paradigm Selection designed to facilitate the development of Morphological Lexicon. Here we propose the concept of *paradigm differentiating measure* (*pdm*) which has been used to map lemmas to Lexical Level Morphological Paradigms.

2 Related Work

Automatic mapping of word to a paradigm have been done earlier for other languages. An n-gram-based model has been developed (Sanchez et al., 2012; Linden and Tuovila, 2009) to select a single paradigm in cases where more than one paradigm generates the same set of word forms. These systems use POS information or some additional user in-

[1] Citation form of words

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 203–208,
Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

put from native language speakers to map words to paradigms, instead of a corpus alone.

Lexicon acquisition methods (Carlos et al., 2009; Clement et al., 2004; Forsberg et al., 2006; Mohammed et al., 2012) exist for many languages that extract lemmas from a corpus and map them to morphological paradigms. Functional Morphology has been used to define morphology for languages like Swedish and Finnish, and tools based on Functional Morphology, namely Extract (Forsberg et al., 2006) which suggest new words for a lexicon and map them to paradigms, have been developed. To be able to use a tool like Extract, the morphology of the language has to be fitted into the Functional Morphology definition.

3 Terminology and Notations Used

Definition (Root, Stem, Base, Prefix and Suffix): A *Root* is the basic part of a lexeme[2] which cannot be further analyzed, using either inflectional or derivational morphology. *Root* is that part of word-form that remains when all derivational and inflectional affixes have been removed. A *Stem* is that part of the word form that remains when inflectional suffixes have been removed. A *Base* (b_i) is that part of the word form to which affixes of any kind can be added. It is a generic term which could refer to a *Root* or a *Stem.* A *Prefix* is a bound morpheme that is attached at the beginning of a *Base.* A *Suffix* $s_i \in \sum^*$ is a bound morpheme that is attached at the end of a *Base.*

Definition (Rule) An ordered 3-tuple (α, β, γ) is said to be a *Rule* used to convert a string x_i to a string y_i where α=”ADD/DELETE” is an operation performed on input string x_i; β=position at which the operation specified in α is to be performed on string x_i; γ=z_i is the argument for the operation to be performed.

Example: If x_i=धांवप*(dhaa.Nvapa)*[3] and *Rule*= (”DELETE”, ”END”, ”प(*pa*)”) where α= ”DELETE”; β=”END”; γ=”प(*pa*)” with respect to above Definition y_i=धांव(*dhaa.Nva*).

Definition (Base Formation Rule (BFR)): An ordered n-tuple of *Rules* which is used to convert lemma l_i to base b_i is said to be a *Base Formation Rule BFR.*

Example: If l_i=भास*(bhasa)*[4] and *BFR*= ((”DELETE”, ”END”, ”स(sa)”),(”ADD”, ”END”, ”श(sha)”)) with respect to above Definition b_i=भाश*(bhasha).*

Definition (Morphological Paradigm): An ordered tuple (ϕ, {(ψ_1, ω_1,γ_1),...,(ψ_n, ω_n, γ_n)}) where

- ϕ=p_i, a unique identifier for the i^{th} paradigm,
- ψ_j=BFR the Base Formation Rule corresponding to the j^{th} Base,
- ω_j = S_k a set of (suffix[5], grammatical feature) ordered pairs corresponding to the j^{th} Base and
- γ_j = A boolean flag which is set to 1 if corresponding suffixes uniquely identify the paradigm i.e. corresponding (ψ_j, ω_j) form the *paradigm differentiating measure*[6].
- n is the total number distinct bases for the paradigm,

is said to be a *Morphological Paradigm* which is used to generate the Inflectional Set i.e. all the inflectional word forms, for the input lemma.

Example: When the paradigm is given by

- ϕ=$P11$,
- ψ_1= ((”DELETE”, ”END”, ”स(sa)”),(”ADD”, ”END”, ”श(sha)”)) is the BFR corresponding to the first Base.
- ω_1 = {(◌े(e), singular oblique case), (◌ेक(eka), singular oblique accusative case), (◌ेकूच (ekaUch), singular oblique accusative case with emphatic clitic), ... }.
- γ_1 = 1
- ψ_2= (”DELETE”, ”END”, ”$\emptyset$”) is the BFR corresponding to the second Base.

[2] Lexeme is the basic unit of meaning. It is an abstract unit of morphological analysis in linguistics, that roughly corresponds to a set of forms taken by a single word

[3] धांवप(*dhaa.Nvapa*) (*to run*)

[4] भास(*bhasa*) (*language*)

[5] Here suffix could be made up of more than one suffix concatenated with each other

[6] *paradigm differentiating measure* has been defined in the following subsection.

- ω_2 = {(◌ो(o), plural direct case), (◌ोच (och), plural direct case with emphatic clitic), ...}

- $\gamma_2 = 0$

- n=2,

If the input lemma = भास(bhaasa), then the first Base is भाश(bhaasha) and the second Base is भास(bhaasa). The word forms generated by the above paradigm are as follows: {भाशे (bhaashe), भाशेक (bhaasheka), भाशेकूच (bhaashekaUch), ... भासो (bhaaso), भासोच (bhaasoch), ...}

Definition (Inflectional Set) A set $W_{p_i l_j}$ of all possible word forms generated by a Morphological Paradigm with p_i as paradigm identifier, for a lemma l_j is said to be the **Inflectional Set** for lemma l_j with respect to paradigm p_i.

Example: If p_i=$P10$, a verb Morphological Paradigm and l_j=*walk* with respect to above Definition $W_{p_i l_j}$={*walk, walks, walking, walked*}.

4 Types of Morphological Paradigms:

A Morphological Paradigm is used to generate the inflectional word forms for a given input lemma. At the Surface Level, a Morphological Paradigm generates a set of word forms which can be expressed in an abstract manner as $\{b_i.s_j$: where b_i is the Base; s_j is the Suffix}. At the Lexical Level, a Morphological Paradigm generates a set of word forms which can be expressed in an abstract manner as $\{l_i$+grammatical features : where l_i is the lemma}.

Example: If the input lemma l_i=*dance*, Word forms generated at Surface Level are {*dancing, danced, dances, ...*} where b_i=*danc*. Word forms generated at Lexical Level are {*dance + present continuous, dance + past perfect, dance + present, ...*}.

Morphological Paradigms can differ from each other either at the Surface Level or at the Lexical Level

Surface Level difference between Morphological Paradigms: Two Morphological Paradigms are said to differ at surface level when they generate different set of word forms at the Surface Level for a given input lemma. Surface level difference implies that at least one of the following two conditions is true.

- ∃ at least one BFR that is not the same amongst them.

- ∃ at least one suffix which is not the same amongst them.

Lexical Level difference between Morphological Paradigms: Two distinct Morphological Paradigms are said to differ at lexical level when they generate same set of word forms at the Surface Level. Lexical level difference implies the following condition is true

- ∃ at least one word form which has different grammatical features in the two paradigms.

Each Morphological Paradigm is unique either at the Surface or Lexical level. We refer to the feature which makes the Morphological Paradigm unique as *paradigm differentiating measure* and is defined as follows

Definition (Paradigm Differentiating Measure) The ordered tuple (ψ_j, ω_j) with respect to Morphological Paradigm Definition above is called *paradigm differentiating measure* if it occurs only once across all possible paradigms.

Example 1: If set A and B represent two sets of word forms generated by two different paradigms p_1 and p_2 respectively which differ at the surface level, for a given lemma. Let set A and B be given as follows:

A= { ($b_1.s_1$,f_1), ($b_1.s_2$,f_2), ($b_1.s_3$,f_3), ($b_1.s_4$,f_4), ($b_1.s_5$,f_5)}

B= { ($b_1.s_1$,f_1), ($b_1.s_6$,f_2), ($b_1.s_3$,f_3), ($b_1.s_4$,f_4), ($b_1.s_5$,f_5)}

where b_j is a base obtained using ψ_j, s_j is the suffix obtained using ω_j and f_j is the corresponding grammatical feature.

From set A and B we observe that the word forms differ only at the second entry namely ($b_1.s_2$,f_2) ∈ A and ($b_1.s_6$,f_2) ∈ B hence the corresponding (ψ_1, ω_2) in p_1 and (ψ_1, ω_2) in p_2 are the *paradigm differentiating measure.*

Example 2: If set C and D represent two sets of word forms generated by two different paradigms p_1 and p_2 respectively which differ only at the lexical level, for a given lemma. Let set C and D be given as follows:

C= { $(b_1.s_1,f_1)$, $(b_1.s_1,f_2)$, $(b_1.s_3,f_3)$, $(b_1.s_4,f_4)$, $(b_1.s_5,f_5)$}
D= { $(b_1.s_1,f_1)$, $(b_1.s_3,f_2)$, $(b_1.s_3,f_3)$, $(b_1.s_4,f_4)$, $(b_1.s_5,f_5)$}
where b_j is a base obtained using ψ_j, s_j is the suffix obtained using ω_j and f_j is the corresponding grammatical feature.
From set C and D we observe that the word forms are same at surface level but corresponding grammatical features differ only at the second entry namely $(b_1.s_1,f_2) \in$ A and $(b_1.s_3,f_2) \in$ B hence the corresponding (ψ_1, ω_2) in p_1 and (ψ_1, ω_2) in p_2 are the *paradigm differentiating measure.*

5 Lexical Level Morphological Paradigm Selection for Konkani Nouns

A Konkani noun lemma can be mapped to more than one Morphological Paradigm. The noun Morphological Paradigms are such that they all differ from each other either at the surface level or at the lexical level. It is not possible to implement a Rule Based System to map noun lemmas to Morphological Paradigms due to ambiguity in paradigm selection presented next.

5.1 Ambiguity in Paradigm Selection for Konkani Nouns

Ambiguity in Paradigm Selection for Konkani Nouns exists due to the following reasons
1. Formative Suffix attachment: There is no known linguistic rule[7] to decide which Formative Suffix is to be attached to the Base to obtain the Inflectional Set. This gives rise to ambiguity in choosing the appropriate paradigm.
Example: When noun lemma does not end with a vowel as in case of the noun lemma पाल(*paala*)(*lizard*); then three possible formative suffixes could be attached which gives rise to three possible Stems namely पाला, पाली, पाले (*paalaa, paalI, paale*). Amongst these three possible Stems only पाली (*paalI*) is the correct choice. However no linguistic rule can be used to arrive at the correct stem thus causing an ambiguity in choosing a correct paradigm for the input noun lemma.

2. Multiple paradigm for single noun lemma: A single noun lemma could be mapped to more than one noun paradigm. This gives rise to another ambiguity is paradigm selection.

Example: For noun lemma मराठी(*maraThI*)(*marathi language or marathi speaking person*); the same lemma will map to two different paradigms for the two different senses namely *marathi language* and *marathi speaking person.* In such a case simply computing Suffix Evidence Value *SEV* is not enough to resolve ambiguity.

3. Lexical level differences in paradigms: Some paradigm differ only at lexical level and generate the same Inflectional Set at surface level. This is another ambiguity challenge faced for paradigm selection.

Example: For noun lemma पान(*paana*)(*leaf*); the same lemma will map to two different paradigms which are same at the surface level. This is because a single form in such paradigm have two different grammatical features as in case of पाना(*paanaa*) which could be *singular oblique form* or *direct plural form* which is a type of ambiguity.

5.2 Problem Statement

Given a set of noun lemmas $LX_N = \{l_i :$ $i = 1$ to n, where n is number of lemmas which map to same surface level morphological paradigms$\}$; a set of Lexical Noun Paradigm List P_LN_L= $\{(p_i, \{(BFR_j, s_l,g_l, pdm_l)\}) : p_i$ is the paradigm identifier, BFR_j is the Base Formation Rule, s_l is the stem formative suffix corresponding to the l^{th} suffix group, g_l is the group identifier corresponding to the l^{th} suffix group, pdm_l is the *paradigm differentiating measure* flag corresponding to the l^{th} stem formative suffix, $i = 1$ to q, where q is number of noun paradigms in L, $j = 1$ to r, where r is number of Bases corresponding to the i^{th} noun paradigm and $l = 1$ to s, where s is number of Noun Suffix Groups corresponding to the j^{th} Base of i_{th} noun paradigms$\}$ and Lexical Training Data Set generate Lexical Level Noun Morphological Lexicon set LX_{NM} =$\{(l_i,p_j) :$ $l_i \in LX_N$, and $p_j \in P_LN_L\}$

[7] Linguistic rule based on noun lemma ending characters alone in absence of knowledge of nouns grammatical gender

5.3 Design of Lexical Level Noun Morphological Paradigm Selection

A training data set is prepared for each Lexical Level Paradigm. The features used in the data set are listed in Table 1.

Table 1: Data Set Features for Lexical Level Paradigm Selection.

Name	Feature Description
PID	The paradigm identifier.
FreqDSF	Number of times the direct singular form of the noun occurs in the corpus
FreqSOF	Number of times the oblique singular form of the noun occurs in the corpus.
FreqPOF	Number of times the oblique plural form of the noun occurs in the corpus.

These features were chosen after observing that, in Konkani Lexical Level Paradigms, for one paradigm, the Direct Singular Form (DSF) and Direct Plural Form (DPF) are the same while for the other paradigm, Direct Plural Form (DPF) and Plural Oblique Form (POF) were the same. In general, these features correspond to those word forms that have multiple grammatical roles i.e. those word forms which cause ambiguity. The intuition behind choosing these features was that, if in one paradigm a particular word form has multiple grammatical roles, than its corresponding relative frequency should differ from the other paradigm where it has a single grammatical role.

Example: Let p_i and p_j be two paradigms which are same at surface level but differ at lexical level. Let l_i be the input lemma. In paradigm p_i, let the word form w_i have two grammatical roles as in case of Konkani word फातर(*phaatara*) (*stone*) which is both Direct Singular Form (DSF) and Direct Plural Form (DPF). In paradigm p_j, let the same word form w_i have only one grammatical role which is Direct Singular Form (DSF) and has a different form w_j for Direct Plural Form (DPF) which is also Plural Oblique Form (POF). Thus in the data set for paradigm p_i, frequency of DSF and POF will follow a different pattern when compared to frequency of DSF and POF in p_j.

To select appropriate machine learning model for the training data set various machine learning algorithms were tested on the training data set. The best performing model namely Logistic Regression was chosen as the learning model as it works well on numeric data, is simple and performed better than other machine learning classifiers as illustrated in Table 2. We created a training data set with 356 noun lemmas and assigned the paradigm identifier manually. This was used as a training model to pick lexical level paradigm for the input lemma. The algorithm for the Lexical Level Morphological Paradigm Selection is illustrated in Figure 1.

Algorithm: Lexical Level Morphological Paradigm Selection

Input: Noun lemma l_i, Lexical Training Data Set TDS, set of unique corpus words W_C, Lexical Noun Paradigm List (P_LN_L), Pruned Relevant paradigm set R_P, Surface Noun Paradigm List (P_LN_S)

Output: Relevant paradigm set with lexical paradigms R_P.

```
/* Select appropriate Lexical Level Paradigm */
For each $p_i \in R_P$
    If $p_i \in P_LN_L$
        /* Compute corresponding Feature Set $FS$ for Lexical Level Paradigm*/
        $FS$ = computeFeatureSet($l_i$,$W_C$, $p_i$, $P_LN_S$)
        $R_{p_i}$ = applyLogisticRegression($TDS$,$FS$)
        Replace $p_i$ with $R_{p_i}$ in $R_P$
    End If
End For
```

Figure 1: Algorithm: Lexical Level Morphological Paradigm Selection for Konkani Noun.

6 Experimental Results and Evaluation

The goal of the experiment was to identify a machine learning model to automatically assign lexical level morphological paradigms to noun lemmas. To choose the model for lexical level paradigm assignment, we ran various classification algorithms on our development data sets created with features listed in Table 1 using 10 fold cross validation to determine the best training model. The performance of machine learning classifiers on our data set are tabulated in Table 2. Here Precision, Recall and F-score are the weighted average values generated.

Table 2: Model Selection for Lexical Level Paradigm Selection.

Algorithm	**Precision**	**Recall**	**F-Score**
Bayesian Classifiers			
Naive Bayes	0.796	0.815	0.785
Bayes Net	0.787	0.806	0.79
Function Classifiers			
Logistic	0.94	0.941	0.94
Multilayer-Perceptron	0.821	0.834	0.822
RBFNetwork	0.806	0.82	0.79
SimpleLogistic	0.958	0.958	**0.957**
SMO	0.839	0.798	0.723
Instance-Based Classifiers			
B1	0.84	0.846	0.842
KStar	0.828	0.834	0.807
Ensemble Classifiers			
AdaBoost	0.915	0.916	0.912
Bagging	0.937	0.938	0.938
Random Sub Space	0.898	0.896	0.887
Decorate	0.952	0.952	0.951
Logit Boost	0.932	0.933	0.93
Rule-Based Classifiers			
PART Decision List	0.94	0.941	0.94
Ridor	0.94	0.941	0.94
ZeroR	0.61	0.781	0.685
Decision Tree Classifiers			
Random Forest	0.928	0.93	0.928
Logistic Model Tree	0.977	0.978	**0.977**
REPTree	0.936	0.935	0.936

Analyzing the performance of the various classifiers from Table 2, We observe that Logistic Regression based models namely SimpleLogistic and Logistic Model Tree outperform other models. Hence Logistic Regression was chosen as a training model to select relevant lexical level morphological paradigm.

7 Conclusion

In this paper we present a method to automatically select a lexical level morphological paradigm for a Konkani noun lemma. We define *paradigm differentiating measure* and use the same to select features and prepare the training data set. The data set thus created in used to identify logistic regression as an appropriate model to select lexical level morphological paradigms for Konkani nouns with an F-score of 0.957.

References

Carlos Sujay Cohan, Choudhury Monojit and Dandapat Sandipan. 2009. *Large-Coverage Root Lexicon Extraction for Hindi.* Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece.

Clement Lionel, Sagot Benoit and Lang Bernard. 2004. *Morphology Based Automatic Acquisition of Large-coverage Lexica.* Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal.

Markus Forsberg, Harald Hammarström and Aarne Ranta. 2006. *Morphological Lexicon Extraction from Raw Text Data.* Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland.

Lindén Krister and Tuovila Jussi. 2009. *Corpus-based paradigm selection for morphological entries.* Proceedings of NODALIDA 2009.

Attia Mohammed, Samih Younes, Shaalan Khaled and Genabith Josef. 2012. *The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words.* Proceedings of COLING 2012, Mumbai, India.

Vícor M. Sánchez-Cartagena, Miquel Esplá-Gomis, Felipe Sánchez-Martínez and Juan Antonio Pérez-Ortiz. 2012. *Choosing the correct paradigm for unknown words in rule-based machine translation systems.* Proceedings of the Third International Workshop on Cambridge University Free/Open-Source Rule-Based Machine Translation, Gothenburg, Sweden.