ISSN No:-2456-2165

Logistic Regression for Lexicon and Evaluation of Morphological Paradigm Selection for Konkani

Alice D Souza DM (Sr.Sadhana DM)
Research Scholar
St Raymond's PU College
Vamanjoor, Mangaluru

Dr. Rio D Souza
Professor, Department of CSE
St. Joseph Engineering College
Mangaluru

Abstract:- We discover the idea of constructing a classifier that can be used to recognize the emotion speech exists in web treatises such as forums and web blogs. The emotions are existing in speech, that emotion can be abstracted into thematic areas of race, religion and nationality. We are using sentimental analysis methods to create a model classifier and specific subjectivity detection is not only used to identify that the given sentences are subjective or not but also used for rate and recognize the polarity of sentiment expressions. Initially we need to removing objective sentences this will leads to the whittling down the size of the document. For speech recognition with the usage of sematic features and subjectivity related to emotion speech we are building a lexicon that is active to build a classifier. The emotion corpus experiments indicate the major practical application for areal-world web discourse.

I. INTRODUCTION

Via conversational speech interface user regularly cooperate with natural language understanding systems. The systems like Microsoft Cortana, Google Now and Apple Siri etc. are all depends on spoken language understanding(SLU.) Construction of these type of systems is challenging because of conversational speech naturally encircles repetitions, spontaneous, partial words, disfluencies and out of vocabulary (OOV) words (De Mori et al.,2008; Huang et al., 2001). In addition to that, for transcription errors spoken language understanding(SLU) system must be robust. These transcription error depends on the domain and task.

Our system consists of some of inputs like, morphological rules, raw text corpus and root word lexicon for Konkani. The output of the system is Noun paradigm repository. Here a noun in the WordNet is assigned its inflectional paradigm. For Konkani nouns we are implementing a one of method automatic paradigm selector. For Konkani nouns we are using our own paradigm structure, that customs Finite State based sequencing of morphemes. This involvement leads to study the pronunciation lexicon from some speech samples. Instead of only focused on either of pronunciation and spelling aspect of lexical entries, we recommend new method that evaluates both the spelling and the pronunciation together. To discovering an optimal lexical illustration of data set we want to move away from a statistic recognition lexicon and discover automatic methodologies.

For learning lexical entries, we need to identify multiple acoustic multiple sample of same phrase or word, automatically proposing both pronunciation and spelling [10]. To overcome the problem of OOV, we need to indexed a database of audio documents, by learning the words present in database.

II. RELATED WORK

Earlier for other languages we have been done automatic mapping of word to a paradigm. To map words to the paradigms (Sanchez, 2012) we are using Rule based system. With the usage of POS information or some additional input from native language speakers we can map the words to the paradigms. Corpus assisted approach (Desai et al.,2012) has been used to map the Konkani verbs to a single paradigm. Lexicon acquisition methods are present in many languages that can be used to map words to morphological paradigms. To describe the morphology for languages like Finnish and Swedish, and tools based on Functional Morphology, like extract by the use of functional morphology, which suggest new words for a lexicon and map them to paradigms. Based on tool Functional Morphology uses constraint grammars to map words correctly to paradigms. The tool extract can be used to fit the morphology of the language into the Functional morphology definition.

III. KONKANI NOUN MORPHOLOGY

Noun forms are usually achieved by attaching prefix or suffixes. Prefixing is not more productive in Konkani language. The suffixes could be inflectional suffixes or derivational suffixes. Typical lexicons preserve the derivational form of word. Later we do not need map derivational words to paradigms. Inflectional forms are not preserved in lexicons hence there is a need to group all inflectional forms into paradigms. Inflectional paradigm for a noun in Konkani will be a set of suffixes, that can be attached to a common noun

A. Konkani Noun Inflectional Morphology

Noun present in Konkani are inflected for syntactic case and number. The number can be a plural or singular. The Konkani cases are dative, locative, instrumental, accusative and genitive. In addition, nominatives, other cases display a suffix before the case marker nominatives, other cases show a suffix before the case marker. Two basic types of Konkani cases direct and oblique (Almeida, 1989). The cases that require oblique suffixes are called oblique case type.

B. Lexicon Building

A number of earlier works that have been developed to generate sentiment words representing positive orientation and negative orientation. There are two main method for generating dictionary, opinion lexicon and corpus-based method. The former comprises static dictionary of semantically relevant words which can be tagged with both reliability and polarity label [17-18]. A number of proposed method generated by the usage of bootstrapping strategy, that uses an online dictionary like WordNet [8] and SentiWordNet [19] and small set of seed opinion words. Substantial resources of dictionaries of opinion lexicon are exist that build mainly from adjectives, and also from adverbs, verbs and nouns [17,20]

Dictionary based method suffer to determine opinion words with context specific orientations and domain. Domain corpus method is used to capture opinion words with a desired syntactic or co-occurrence forms. To determine the semantic orientation of words we are using natural language processing rule-based techniques, syntactic, structural and sentence level features. A lexicon is populated with phrases and words, which are more attuned to the field by incorporating contextual features, sematic orientation of an opinion word can be altered theoretically. To reduce or amplify the intensity of a neighboring lexicon item we are using Intensifier feature. The [18] phrase-level sentiment analysis method decides whether the expression is polar or neutral and then disambiguates the polarity of the polar expression.

C. Terminology and Notations Used Definition (Root, Stem, Base, Prefix and Suffix):

The basic part of lexeme is Root. That cannot be further examined, by using either derivational morphology or inflectional morphology. Root is the part of word-form when all derivational affixes have been removed. Stem is the part of the word-form when inflectional suffixes have been removed. A Base(bi) is the part of word-form to which affixes of any type can be added. The Prefix is a bound morpheme which is attached at the beginning of a Base Suffix si ∈ Σ*is a bound morpheme which is attached at the end of the Base.

➢ *Definition (Rule)*

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

- Example: If xi= (dhaa.Nvapa) and Rule= (” DELETE”,” END”,” (pa)”) where α=” DELETE”; β=” END”;γ =” (pa)” with respect to above Definition yi= (dhaa.Nva).

➢ *Definition (Base Formation Rule (BFR)):*

An ordered n-tuple of Rules. Which is used to convert lemma li to base bi is said to be a Base Formation Rule (BFR)

- Example: If li= (bhasa) and BFR= ((”DELETE”,”END”,” (sa)”),(”ADD”, ”END”, ” (sha)”)) w.r.t above Definition bi= (bhasha).

➢ *Definition (Morphological Paradigm):*

An ordered tuple (ϕ, {(ψ1, ω1, γ1),...,(ψn, ωn, γ n)}) where

- ✓ ϕ=pi, a unique identifier for the ith paradigm,
- ✓ ψj=BFR the Base Formation Rule corresponding to the jth Base,
- ✓ ωj = Sk a set of (suffix, grammatical feature) ordered pairs corresponding to the jth Base and
- ✓ γj= A Boolean flag which is set to 1 if corresponding suffixes uniquely identify the paradigm i.e. corresponding (ψj, ωj) form paradigm differentiating measure.
- ✓ n - total number distinct bases for the paradigm,

The Morphological Paradigm is used to generate the Inflectional Set i.e. all the inflectional word forms, for the input lemma.

- Example: The paradigm is given by
- ✓ ϕ=P 11,
- ✓ ψ1= ((”DELETE”, ”END”, ” (sa)”),(”ADD”, ”END”, ” (sha)”)) - BFR equivalent to first Base.
- ✓ 1 = {(ॅ (e), singular oblique case), (ॅ (eka), singular oblique accusative case), (ॅ (eka Uch), singular oblique accusative case with emphatic clitic), ... }.
- ✓ γ1 = 1
- ✓ ψ2= (”DELETE”, ”END”, ”ȼ”) - The BFR equivalent to the second Base. ω2 = {((o), plural direct case), ((och), plural direct case with emphatic clitic), ...}
- ✓ γ2 = 0
- ✓ n=2,

If the input lemma = (bhaasa), then the first Base is (bhaasha) and the second Base is (bhaasa). The word forms generated by the above paradigm are as fol-lows: { (bhaashe), (bhaasheka), - (bhaashekaUch), ... (bhaaso), (bhaasoch), ...}

➢ *Definition (Inflectional Set)*

A set Wpilj of all possible word forms produced by a Morphological Paradigm with pi as paradigm identifier, for a lemma lj is said by the In-flections Set for lemma lj with respect to paradigm pi.

- Example: If pi=P 10, a verb Morphological Paradigm and lj=walk with respect to above Definition Wpilj ={walk, walks, walking, walked}.

D. Types of Morphological Paradigms:

For a given input lemma, we need to generate the inflectional word forms by using Morphological Paradigm. At the Surface Level, Morphological Paradigm produces a set of word forms which can be expressed in an abstract method as {bi.sj: where bi is the Base; sj is the Suffix}. At the Lexical Level, a Morphological Paradigm produces set of word forms that can be expressed in an abstract method as {li+ grammatical features: Here, li is the lemma}.

- Example: If the input lemma li=dance, At the surface level word forms are produced, i.e. {warning, warned, warns...} where bi=warn. At the lexical level generated Word forms are {warn + present continuous, warn + past perfect, warn + present.}. Morphological Paradigms can be vary from each other either at lexical level or at the surface level.

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

- 9 at least one BFR that is not the same amongst them.
- 9 at least one suffix which is not the same amongst them.

Lexical Level difference between Morphological Paradigms: At the lexical level the two distinct Morphological Paradigms are said differ when they produces same set of word forms at the surface level. The following condition is true for lexical difference implies

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

➢ *Definition (Paradigm Differentiating Measure)*

The ordered tuple (ψj, ωj) w.r.t. Morphological Paradigm Definition above is named paradigm differentiating measure if it happens only once across all potential paradigms.

- Example 1: Two set of word forms Set A and Set B that can be generated by two different paradigms p1 and p2 respectively which differ at surface level, for given input lemma(li). Set A and B is given by

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), (b1.s4,f4), (b1.s5,f5)}

Where,
bj –base obtained using ψj,
sj - suffix obtained using ωj and
fj - corresponding grammatical.

We observe that from set A and B the word forms differ only at the second entry namely (b1.s2,f2) 2 A and (b1.s6,f2) ∈ B hence the corresponding (ψ1, ω2) in p1 and (ψ1, ω2) in p2 are the paradigm differentiating measure.

E. Lexical Level Morphological Paradigm Selection for Konkani Nouns

Konkani noun lemma are mapped to the more than one Morphological Paradigm. The noun Morphological Paradigms are differing from each other either at the lexical level or at surface level. Due to the ambiguity in paradigm selection present in next it is not possible to implement a Rule Based System to map noun lemmas to Morphological Paradigms.

Ambiguity in Paradigm Selection for Konkani Nouns Due to following reasons Ambiguity in Paradigm Selection for Konkani Nouns exists

➢ *Formative Suffix Attachment:*

To obtain the Inflectional Set there is no known linguistic method to decide which Formative Suffix is to be attached to the Base. This leads to ambiguity in choosing the appropriate paradigm

- Example: As in case of noun lemma noun(paala)(lizard) when lemma does not end with a vowel; then 3 likely form-tiye suffixes could be attached which leads to the three likely Stems namely, (paalaa, paalI, paale). Amongst these 3 possible stems only (paall) is the correct choice. Though linguistic rule cannot be used to arrive at

the correct stem thus causing an ambiguity in choosing a correct paradigm for the input noun lemma.

F. Lexical Level Differences in Paradigms

At lexical level Some paradigm differs and produces the same Inflectional Set at surface level. This is the another ambiguity challenge which is faced for paradigm selection.

- Example: For noun lemma (paanaa) (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 has two different grammatical features as in case of (paanaa) that could be direct oblique form or singular form which is type of ambiguity.

- Algorithm:Lexical Level Morphological Paradigm Selection

➢ Input: Lexical Training Data Set TDS, set off unique corpus words WC Lexical Training Data Set TDS, set of unique corpus words WC, Noun lemma li, Lexical Noun Paradigm List (PLNL), Pruned Relevant paradigm set RP, Surface Noun Paradigm List(PLNS).

➢ Output: Relevant paradigms set with lexical paradigms RP.

```
/* Select appropriate Lexical Level Paradigm */
For each pi 2 RP
  If pi 2 PLNL

    /* Compute corresponding Feature Set FS for Lexical Level Paradigm*/
    FS = compute Feature Set (li,WC, pi, PLNS)

    Rbi = apply Logistic Regression (TDS,FS) Replace pi with Rpi in Rp

  End If End For
```

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

IV. EXPERIMENTAL RESULTS AND EVALUATION

The main goal of the experiment was to recognize a machine learning model to automatically as-sign lexical level morphological paradigms to noun lemmas. To select the model for lexical level paradigm assignment, we competed numerous classification algorithms, our development data sets created with features recorded in Table 1 using 10-fold cross validation to determine the best training model. The presentation of machine learning classifiers on our set is tabulated in Table 2.

Algorithm	Precision	Recall	F-Score
Bayesian Classifier			
Naive Bayes	0.796	0.815	0.785
Bayes Net	0.787	0.806	0.79
Function Classifier			
Logistic	0.94	0.941	0.94
Multilayer-Perceptron	0.821	0.834	0.822
RBF Network	0.806	0.82	0.79
Simple logistic	0.958	0.958	**0.957**
SMO	0.839	0.798	0.723
Instance-Base Classifiers			
B1	0.84	0.846	0.842
K Star	0.828	0.834	0.807
Ensemble Classifier			
Ada Boost	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 Classifier			
PART Decision List	0.94	0.941	0.94
Ridor	0.94	0.941	0.94
Zero R	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**
REF Tree	0.936	0.935	0.936

Table 2:- Lexical Level Paradigm Selection Model

From table 2, we are analyzing the performance of the various classifiers, we observe that Logistic Regression based models like Simple Logistic and Logistic Model Tree outperform other models. To select relevant lexical level morphological paradigm, we are chosen Logistic Regression as a training model.

V. CONCLUSION

In this paper we present a method for a Konkani noun lemma to automatically select a lexical level morphological paradigm. We define paradigm differentiating measure and which is used to select prepare the training data set and features. To select lexical level morphological paradigms for Konkani nouns with an F-score 0.957 we are using created data set to identify logistic regression as an appropriate model.

REFERENCES

[1]. H. Bourlard and N. Morgan. 1993. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Norwell, MA.
[2]. G. E. Dahl, D. Yu, and L. Deng. 2011. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In Proc. ICASSP.
[3]. G. E. Dahl, T. N. Sainath, and G. E. Hinton. 2013. Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout. In ICASSP.
[4]. R. De Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur. 2008. Spoken language understanding. Signal Processing Magazine, IEEE, 25(3):50–58.
[5]. X. Glorot, A. Bordes, and Y Bengio. 2011. Deep Sparse Rectifier Networks. In AISTATS, pages 315– 323.
[6]. S. Goldwater, D. Jurafsky, and C. Manning. 2010.Which Words are Hard to Recognize? Prosodic, Lexical, and Disfluency Factors That Increase Speech Recognition Error Rates. Speech Communications,52:181–200.
[7]. A. Graves and N. Jaitly. 2014. Towards End-to-End Speech Recognition with Recurrent Neural Networks.In ICML.
[8]. A Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber.2006. Connectionist temporal classification:Labelling unsegmented sequence data with recurrent neural networks. In ICML, pages 369–376. ACM.
[9]. K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn.2013. Scalable modified Kneser-Ney language model estimation. In ACL-HLT, pages 690–696, Sofia, Bulgaria.
[10]. G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed,N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen,T. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speechrecognition. IEEE Signal Processing Magazine,29(November):82–97.
[11]. 11.X. Huang, A. Acero, H.-W. Hon, et al. 2001. Spoken language processing, volume 18. Prentice Hall Engle wood Cliffs.
[12]. A. Maas, A. Hannun, and A. Ng. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In ICML Workshop on Deep Learning for Audio, Speech, and Language Processing.
[13]. T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH, pages 1045–1048.
[14]. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek,K. Vesel´y, N. Goel, M. Hannemann, P. Motlicek,
[15]. Y. Qian, P. Schwarz, J. Silovsky, and G. Stemmer. 2011. The kaldi speech recognition toolkit. In ASRU.T. N. Sainath, I. Chung, B. Ramabhadran, M. Picheny,J. Gunnels, B. Kingsbury, G. Saon, V. Austel, and U. Chaudhari. 2014.
[16]. Parallel deep neural network training for lvcsr tasks using blue gene/q. In INTERSPEECH.G. Saon and J. Chien. 2012.
[17]. Large-vocabulary continuous speech recognition systems: A look at some recent advances. IEEE Signal Processing Magazine, 29(6):18–33.
[18]. I.Sutskever, J. Martens, and G. E. Hinton. 2011. Generating text with recurrent neural networks. In ICML,pages 1017–1024.
[19]. I. Sutskever, J. Martens, G. Dahl, and G. Hinton. 2013.On the Importance of Momentum and Initialization in Deep Learning. In ICML. K. Vesely, A. Ghoshal, L. Burget, and D. Povey. 2013.
[20]. Sequence-discriminative training of deep neural networks.In Interspeech M. D. Zeiler, M. anzato, R. Monga, M. Mao, K. Yang,Q.V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean and G. E. Hinton. 2013.
[21]. Rectified Linear Units for Speech Processing. In ICASSP.