CONTEXT AND LEXICON BASED GENDER IDENTIFICATION OF NOUN PHRASE FOR GUJARATI TEXT USING HYBRID APPROACH

Ms. Chetana Tailor, Dr. Bankim Patel

Abstract -- To understand a language, analysis shall be performed at word level, sentence level, semantic level and discourse level. Morphological analysis is a foundation to perceive the meaning of a sentence. One of the essential tasks at morphological analysis is gender identification of a noun phrase that plays a pivotal role in Natural Language Processing Applications like Anaphora Resolution, Machine Translation and Text Summarization. It is a difficult task as there are special cases in Gujarati nouns that are contrasted with regular Gujarati linguistics rules for identifying the gender of the noun phrase. We have proposed solution for context based gender identification of noun phrase using Hybrid approach incorporating lexicon comprises of Gujarati nouns. Achieved accuracy of context based gender identification of noun phrase is 98.63% on development set and 88.62% on testing corpus.

Index Terms--Natural Language Processing, Morph Analysis, Gujarati

I. INTRODUCTION

Gender identification is one of the essential tasks of Morphological analysis. It can be further used in Natural Language Processing Applications like Anaphora Resolution, Machine Translation and Text Summarization. Gender identification for English language is simple as it has simple morphology [3, 7]. Noun, pronoun, adjective and verb are inflected according to the gender of the noun in Gujarati and Hindi language [4,7]. Order of inflections followed in Gujarati noun is: Noun + Gender marker + Number marker + Case marker [4].

Agglutinative nature of Gujarati language increases the complexity of the Gender identification task as inflections are attached with noun itself as shown in the general structure of a noun. Recognition of the gender marker from the Gujarati noun is challenging as compared to Hindi as case markers are not attached with noun in Hindi language. In Gujarati language, there are three categories of Gender: Masculine, Feminine, and Neuter with gender marker ' ો'- o - o, 'ી'-ii-e, and 'ુ'-u-u for respective category [1, 4]. In table 1 we have listed nouns having contradiction with the common gender markers. Gender marker is part of the noun. This two are increasing the complexity of the task to identify the gender of a noun.

Rule based approach used in [5,6,8] alone is inefficient for Gujarati language as there are exceptions in nouns. Few of them are listed in Table 1. Paradigm based approach is useful when inflections in the word are less [2]. Gujarati language is rich in morphology so creation of paradigm for it is tough and time consuming. Though the morphological richness of Gujarati language is useful in statistical machine learning as a contextual feature as it is efficient for Hindi language [7]. The elementary necessity for implementing Statistical approach is a large amount of annotated data. Due to unavailability of Gujarati annotated corpus, task of Gender identification of a noun phrase becomes tough. Therefore, to overcome the limitation of Rule based and Statistical approach, we have adopted the Hybrid approach [11] for Gender identification of a noun phrase for Gujarati text.

II. LITERATURE SURVEY

Table 1: Gender Marker Exceptions for Nouns [11]

Masculine	Feminine	Neuter
પત્ર-pəʈ̪rə-letter	વાર્તા –varta- story	ઘર-gʱərə - home
ટપાલી -ʈ̪əpälii – postman	જળો -dʒəl̪o- leech	ઘી - gʱii-clarified butter
મુખી-mukhii – headman	ડરો-ɖ̪əro - young grass	બી -bii –seed
ઇરુ –iru- snake	આબરુ -abəru –fame	મોં –moä- mouth
શબ્દ- ʃəbɖ̪ə- word	આંખ-aäkʰə - eye	શાક-ʃakə - vegetable
વાળ-väl̪ə -hair	શાન –ʃanə -sense	ફળ-pʰəl̪ə -fruit
મહિમા-məɦima –fame	જમીન -dʒəmiinə -land	કપાળ -kəpal̪ə-forehead
સાધુ -saɖ̪ʱu- hermit	શાળા -ʃal̪a- school	મોતી-moʈ̪ii –pearl
ડાકુ -ɖ̪aku- rogue	ઘો -gʱo- lizard	પાણી-paɳ̪ii –water
ધોબી - ɖ̪ʱobii-washerman	સાસુ-sasu - mother-in-law	દહીં -ɖ̪əɦiiä- curd
સાબુ –sabu- soap	જીભ – dʒiibʱə-tongue	લોહી - loɦii-blood
પાઠ –pathə- lesson		પંખી-pəäkʰii – bird
આત્મા –atma- soul		

Paradigm based[2], Rule based[5, 6, 8], Statistical supervised machine learning approach[7] and Hybrid approach[11] have been implemented for Gender identification for Indian languages.

In [2] Paradigm approach for Hindi morphological analyzer and generator is used. Paradigms are developed based of taxonomy of words containing noun, pronoun, adjective, verb, adverb and excluded the proper noun based on the inflectional patterns found in words. Gender, number and case of the noun is used key features used for designing paradigms for noun class. Accuracy of the system is unpublished. This approach is efficient for languages with less inflections. As we have discussed previously Gujarati is morphologically and inflectionally rich language with three genders and exceptions in gender inflections for nouns. So, designing paradigm for Gujarati text for gender identification is challenging as compared to Hindi language.

In [5], the rule based approach is adopted for morph analysis. Rules according to the lexical category of the word are designed using suffix of a noun which is quite similar to paradigm based approach used in [2]. This work is further enhanced in [8] by them. After replacement of case, number and gender marker, words are searched in the database containing root words. Though they have achieved 87.48% accuracy, limitation of the system is it is highly dependent on the lexicon and exhaustive rules.

Rule-based approach with dictionary for common nouns is adopted for Gujarati language [6]. According to the POS tag of words, rules are made. Though accuracy for gender identification of noun is 92.06%, rules for noun morphology are contradicted with Gujarati grammar found in [1] as well as discussed previously.

In [7], context based morph analysis has been performed by considering four values of a gender namely feminine, masculine, any and none. SVM classifier with word level and sentence level features is used for gender identification. List of word level features are: word, last 2 characters, last 3 characters, last 4 characters, lemma, and word length. List of sentence level features are lexical category, next word, and previous word. On Hindi Treebank" corpus, 96.19% accuracy has been achieved. As scarcity of annotated corpus, this approach is not advisable.

Paradigm based approach[2] is successful for language with less inflections. Context of the word plays a key role which is neither considered in Rule based approach nor in Paradigm based approach [2, 5, 6, 7] which affects the result when word is neither found in lexicon nor matched with any of the designed rules. Rule based approach[5, 6, 8] is unsuitable for inflectionally rich languages as formation of

exhaustive rules for language is time consuming and difficult task. Results of this approach can be improved by combining lexicon with it but in this approach also size and dependency on lexicon are two major factors. In this approach too, if noun is not in the lexicon, performance gets affected [8]. In [7], context based morphology using Statistical approach for gender identification is used for Hindi language. This approach is independent of lexicon but it needs large amount of data for training and testing purpose. Neither exhaustive rule base system nor merely statistical approach for Gujarati language is possible as existence of exceptions to common linguistics rules and unavailability of annotated corpus. Therefore, we have adopted the context based gender identification for Gujarati language in which the context of words in current sentence as well as inflections in adjective and verb are used to improve the performance of the system. Thus we have combined the Statistical supervised machine learning approach and Rule based approach with lexicon. In next section, design and adopted approach to identify the gender of the noun phrase is discussed.

III. METHODOLOGY

According to the literature study as discussed in previous section for Indian languages confirms that Hybrid approach is suitable for gender identification of noun phrase for Gujarati language. In this system, three main components are designed and developed: lexicon containing common nouns, suffix list of personal nouns, supervised statistical model SVM, and linguistics rules. SVM is a binary classifier which has to be extended for multiclass classification with pair-wise classification using YamCha[9] tool as it beats the approach one versus all. For developing the statistical model, considered window size is three. Thus achieved accuracy for statistical model SVM is 63.77% with 52.17% precision, 62.50% recall and 56.87 F1 score on development data set. Following rules[11] are designed to process the data after applying Statistical supervised SVM model and output after applying rules is any one of masculine, feminine, nueter or any:

Rule 1: If noun phrase comprises of proper noun:

Rule 1.a:

If noun phrase contains common noun then noun is searched in lexicon and if noun is found in the lexicon, then gender is assigned according to the lexicon value.

Rule 1.b:

Else first proper noun in the noun phrase is matched with the suffix listed in the lexicon. If first proper noun is matched with any one of the suffix from lexicon containing proper noun suffix listed in lexicon, then gender is assigned according to the matched suffix.

Rule 2:If noun phrase does not contain proper noun and noun phrase contains one or more common nouns, then:

Rule 2.a:

If last two common nouns makes compound noun then gender for that compound noun is search in the lexicon of compound noun. If compound noun is found in the lexicon, then gender for the noun phrase is assigned accordingly.

Rule 2.b:

Else if last noun of the noun phrase is searched in the lexicon, and if noun is found in the lexicon then gender is assigned according to thedictionary value.

Rule 2.c:

Else if noun phrase contains one or more adjective,

Rule 2.c.1:

If adjective has any one of the suffix “ു” , “ൂ”, “ാ" , gender of the noun phrase is neutral.

Rule 2.c.2:

Else if adjective has any one of the suffix “ി" or " “ാ", gender of the noun phrase is masculine.

Rule 2.c.3:

Else if adjective has the suffix “ീ", gender of the noun phrase is feminine.

Rule 2.d:

Rule 2.d.1:

If last noun in noun phrase has suffix “ુ”, “ુ઼”, or “ાઁ”, gender of the noun phrase is neutral.

Rule 2.d.2:

Else if last noun in noun phrase has any one of the suffix “ો", “ઓ" or "ા”, gender of the noun phrase is masculine.

Rule 2.d.3:

Else if last noun in noun phrase has any one of the suffix “ી”, “ઓ", "ઈ" or "ઈઓ", gender of the noun phrase is feminine.

Rule 2.e:

Else predicted value of the SVM model is considered as a gender of the noun phrase.

Rule 3: If noun phrase does not have any common noun and only made up of quantifier, its gender is any.

Rule 4: if noun phrase is made up of demonstrative noun and does not have any common noun than noun phrase is assigned the category “any” as a gender of the noun phrase.

Dataset is discussed in the next section.

IV. DATA SET

We have collected Gujarati news articles to create a corpus for Statistical supervised machine learning. Corpus is annotated by following the guideline defined under the AnnCorra : Annotating Corpora, Guidelines For POS And Chunk Annotation For Indian Languages[10] for POS tagging and chunking as these are used as features for identifying the gender of the noun phrase.

In [11], system has been tested on total 480 tokens with 350 phrases are tagged for training purpose. Among them 61 are feminine noun phrase, 57 are masculine, 30

are neutral noun phrases, and 15 are of any one of these three categories of noun phrases. Testing is done on 949 tokens with 686 phrases. Corpus size has been increased and now system has been tested on total 2050 noun phrases.

For finding average word length, dictionary with 6,953 words[9] is created which is used for selecting features for SVM learning. Average word length of the word is 6 Unicode characters calculated using 6,953 dictionary words for selecting n-gram of word characters. Average word length is used to find the n-gram of the words used as a feature in designing statistical model. POS tag is used as to identify lexical category of the word and which helps to design rules. Chunk tags are used to identify the boundary of the noun phrase. Chunk tags, word without using case markers, and word length are selected as features for statistical model development. The lexicon contains total 4485 common nouns in which 1,743 masculine nouns, 1,652 feminine nouns and 1,090 neutral nouns [9]. In the next section, we have discussed the result analysis and conclusion.

V. RESULT ANALYSIS AND CONCLUSION

Accuracy of Support Vector Machine model is 63.77% due to small amount of training data set. After applying rules on the processed data of model, accuracy of this system is improved upto 98.63% on development set and 88.62% on testing data set. Though Gujarati language has exceptions listed in Table 1, system is still able to identify the correct gender of the noun phrase. A major issue noticed by us is word with Genitive case marker because in few words, Genitive case marker is part of the lexeme and not indicating genitive case. Hybrid approach combining Statistical Machine learning approach, Rules and lexicon performs good compared to existing Gender identification system for Gujarati as context based rules are applied as well as default value is not assigned if noun is not in the lexicon.

VI. REFERENCES

[1] B. Suthar, “Gujarati-English Learner's Dictionary,” 2003. [Online]. Available: http://ccat.sas.upenn.edu/plc/gujarati/guj-engdictionary.pdf. [Accessed: 02-Oct-2017].

[2] V. Goyal and G. S. Lehal, “Hindi Morphological Analyzer and Generator,” *2008 First International Conference on Emerging Trends in Engineering and Technology*, pp. 1156–1159, 2008.

[3] S. Vikram, “Morphology: Indian Languages and European Languages,” *International Journal of Scientific and Research Publications*, vol. 3, no. 6, pp. 1–5, Jun. 2013.

[4] S. Upadhyay, "Nouns in english and gujarati a comparative linguistic study," thesis, 2015.
[5] U. Kapadia, A. Desai, "Morphological Rule Set and Lexicon of Gujarati Grammar: A Linguistics Approach" VNSGU, vol. 4, no. 1, pp. 127-133, Jul 2015.
[6] S. Maurya, et al., "Gender and Number Identification for Gujarati word: Rule-Based Approach," NJSIT, vol. 9, no. 2, pp.1-7, Dec. 2016.
[7] D. K. Malladi, "Context Based Morphological Analysis," Dissertation, IIITH, 2016.
[8] U. Kapadia and A. Desai, "Rule Based Gujarati Morphological Analyzer," International Journal of Computer Science Issues, vol. 14, no. 2, pp. 30–35, Mar. 2017.
[9] T. Kudo and Y. Matsumoto, "YamCha: Yet Another Multipurpose Chunk Annotator." [Online]. Available: http://chasen.org/~taku/software/YamCha/index.html. [Accessed: 20-Jun-2017].
[10] Bharati, et al (2006). AnnCorra : Annotating Corpora, Guidelines For POS And Chunk Annotation For Indian Languages. LTRC-TR31.
[11] Tailor and B. Patel, "Context based Gender Identification of Noun Phrase for Gujarati Text," in *National Conference on Emerging Technologies in IT*, Surat, 2019.

AUTHORS' PROFILE

Ms. Chetana Tailor has more than 9 years of teaching experience and is keen in research work related to Natural Language Processing. She has presented research papers at national and international conference. Her publications include papers related to Gujarati language processing. Her area of interest are Artificial Intelligence, Natural Language Processing and Machine Learning.

Dr. Bankim Patel is leading in academia since 31 years and has been a pioneer in development of Gujarati language computation in field on Computer Science at SRIMCA Bardoli Gujarat. His research work has mainly focused to make a machine understand Gujarati, whether in form of computer language processing, Natural Language Processing or even Digital Image Processing. He has also guided research work and projects for specially-abled persons. He has published and presented various papers at both national and international level. He has guided 13 research scholars and has secured multiple research grants from renowned organization.