Morphology of Dogri Adjectives through Paradigm Structures

Shubhnandan S. Jamwal
PG Department of computer Science and IT, University of Jammu
jamwalsnj@gmail.com

Abstract
The development of the morphological engine for Indian language has always remained a challenge. To develop the morph analyzer the study of the morphology, which shows the internal structure of the words, for morphological rich language remained a challenge. The morphological analysis and generation are essential steps in any NLP application for the development of the different tools. In this paper the morphology of the adjectives in Dogri language has been presented because the adjectives are integral to morphological studies as they highlight how languages encode descriptive and comparative information through word structure and transformation. Their inflectional and derivational properties contribute significantly to the richness of language.

Keywords
Dogri Language, Morphology, Morphological Analyzers, Adjectives, Indian Languages, Paradigm, Indo-Aryan language

1. Introduction

The most significant tool in the development of the NLP application is morphological engine and the adjectives play a significant role in morphology as they contribute to the structure and meaning of words and sentences in a language. In morphology, the study of word formation and structure, adjectives are analyzed for their form, inflection, derivation, and function. In inflectional morphology adjectives often show agreement with the noun they modify in terms of number, gender. In derivational morphology adjectives are often derived from other word classes, such as nouns or verbs, through the addition of suffixes or prefixes. For example: hope (noun) → hopeful (adjective), care (verb) → careful (adjective). Sometimes adjectives can also serve as the base for deriving other forms like adverbs (quick → quickly) or nouns (happy → happiness). Adjectives play a role in compounding and affixation, contributing to the creation of complex words. Adjectives are also helpful in semantic role of the language when they modify nouns, adding descriptive, qualitative, or restrictive meanings. Morphologically, their structure helps convey nuanced meanings.

2. Literature review

M. Tanaka and H. Yamamoto [1] showed that in modern Japanese, emotive adjectives and verbs can be used to express the same emotions, such as kanashii and kanashimu. In the case of emotive adjectives, the agent of emotion is restricted to the first grammatical person. In Heian Japanese, the agent of emotive adjective sentences is limited to the first person, as in Modern Japanese, while the agent of emotive verbs is limited to the second and third person, contrary to Modern Japanese usage. They conclude that the restrictions on choosing between emotive adjectives and verbs in a sentence based on the grammatical person of the agent of emotion are clearer in kanbun-kundoku-tai than in wab un-tai. D. M. Harikrishna and K. S. Rao [2] classified

Hindi stories into three genres: fable, folk-tale and legend. They proposed a framework for story classification using keyword and Part-of-speech (POS) based features. Keyword based features like Term Frequency (TF) and Term Frequency Inverse Document Frequency (TFIDF) are used. Effect of POS tags like Noun, Pronoun, Adjective etc., are analyzed for different story genres. Classification performance is analyzed using different combinations of features with three classifiers; Naive Bayes (NB), k-Nearest Neighbour (KNN) and Support Vector Machine (SVM). V. Goyal and G. S. Lehal [3] presented the morphological analysis and generator tool for Hindi language using paradigm approach for Windows platform having GUI which has been developed as part of the development of a machine translation system from Hindi to Punjabi Language. G. Dhopavkar and M. Kshirsagar [4] presented the work related to syntactic annotation of Marathi text using Ruled-based approach which is very essential in Sense Disambiguation of a natural language text. They implemented a system for generating and applying natural language patterns to overcome the sense ambiguity problem. They manipulate the grammatical structure of sentence to give the correct output for Marathi Language. Some patterns describe the main constituents in the sentence and some, the local context of the each syntactic function. They also discussed the morphological analysis method used for Marathi Language. R. C., D. K., R. Ravindran and K. P. Soman [5] considered incorporating rule based reordering and morphological information for English to Malayalam statistical machine translation. The main ideas which have proven very effective are (i) reordering the English source sentence according to Malayalam syntax, and (ii) using the root suffix separation on both English and Malayalam words. The first one is done by applying simple modified transformation rules on the English parse tree, which is given by the Stanford Dependency Parser. The second one is developed by using a morph analyzer. D. Chakrabarti and P. Bhattacharyya [6] observed that preliminary attempts have been made to class the simple verbs of Hindi into different morphological paradigms along with the phonological changes. At present, the work is limited only to the syntactic structure. Semantic characteristics have not been dealt with. Focus is mainly given on nominal, transitive-causative, and intransitive-causative variants. D. Chakrabarti and P. Bhattacharyya [7] presented the morphological analysis of Sindhi language in which important areas of Sindhi morphemes including structure, function, & nature, categories of words like compound words, prefix words, suffix Words & prefix-suffix words, and writing system are analyzed and reviewed. Moreover, comparative analysis is also carried out to comprehend the formation of Sindhi Morphology. J. Sheth and B. Patel[8] suggested DHIYA a stemmer for Gujarati language. This stemmer is based on the morphology of Gujarati language. To develop the stemmer, inflections which appeared most in Gujarati text were identified. Based on it, the rule set was created. For training and evaluation of the stemmer's performance the EMILLE corpus is used. The accuracy of the stemmer is 92.41%. P. Das and A. Das [9] implemented the finite-state transducer grammar for Linguistic Resource which gives way to the development of Bengali Noun Morphological Analyzer. The final output obtained is around 44% accuracy. This accuracy can be always improved with time if we keep on increasing the nominal roots in the FST grammar file. R. Haque, S. Penkale, J. Jiang and A. Way [10] employed a simple suffix-stripping method for lemmatizing inflected Bengali words and showed that our morphological suffix separation process significantly reduces data sparseness. They also showed that an SMT model trained on suffix-stripped (source) training data significantly outperforms the state-of-the-art phrase-based SMT (PB-SMT) baseline.

3. Dogri Morphology: Adjective

Dogri, an Indo-Aryan language spoken in northern India, exhibits a rich and systematic adjective morphology. Adjectives in Dogri agree with the nouns they modify in gender, number, and sometimes case. Most adjectives are inflected to match the grammatical gender (masculine or feminine), case (direct, oblique and vocative) and number (singular or plural) of the noun. In Dogri, adjectives can be classified based on their agreement with nouns in terms of number, gender, and case. The adjectives play a crucial role in adding detail and description to the language. □□□ □□□□ (Tall boy), □□□ □□□□□ (Tall girl), □□□□□ □□□□ (Cheap phone) and □□□□□ □□□□□ (Cheap saree) are some of the examples of Dogri adjective. These examples illustrate how Dogri adjectives are used to modify nouns and how they can agree with the gender and number of the nouns they describe. The classifications of adjectives in Dogri considering these criteria are as follows:

- **Number Agreement**:
  - **Singular Adjectives**: These adjectives agree in number with singular nouns. For example, "□□□" (Tall) with a singular masculine noun like "□□□□" (boy).
  - **Plural Adjectives**: These adjectives agree in number with plural nouns. For instance, "□□□" (Tall (plural)) with a plural masculine noun like " □□□□ " (boys).
- **Gender Agreement**:
  - **Masculine Adjectives**: Adjectives that agree in gender with masculine nouns. For example, " □□□ " (tall) with masculine nouns like " □□□□ " (boy).
  - **Feminine Adjectives**: Adjectives that agree in gender with feminine nouns. For instance, "□□□" (tall) with feminine nouns like " □□□□ " (girl).
  - **Neutral Adjectives**: Some adjectives may be considered neutral in terms of gender and can be used with nouns of any gender. . For instance, " □□□□□□ " (pink)
- **Case Agreement:**
  - **Direct Case Adjectives:** This is the default case for adjectives, where they typically agree with the noun they modify in terms of gender and number. The direct case is used in various sentence constructions like □□□ □□□□ (Tall boy), □□□ □□□□□ (Tall girl).
  - **Vocative Case Adjective**: In this case, adjectives are used when addressing or calling someone directly. The vocative case is used to get someone's attention or express emotions directly to a person. "□! □□□□ " , "□□□□ □□□□, □□! "
  - **Oblique Case Adjectives:** Adjectives in the oblique case may undergo changes in form, such as adding suffixes or altering their endings. This is typically done to reflect a different grammatical relationship with the noun like "□□□□ □□□□ □□ □□□□ "

4. Proposed Methodology

The paradigm approach to the morphological analysis of Dogri adjectives involves a structured and systematic process to understand their inflectional and derivational patterns, with a focus on how these patterns vary across grammatical categories like gender, number, and case. This methodology begins with the creation of an adjective dataset. The collected adjectives are then analyzed for their root forms (stems) and the affixes or modifications they undergo when inflected to agree with nouns in gender (masculine, feminine), number (singular, plural), and sometimes case (direct, vocative and oblique). Adjectives are categorized into paradigms based on shared morphological behaviors. For instance, adjectives that take the same inflectional pattern are grouped into one paradigm, while others with different inflectional markers form distinct paradigms. After the creation of paradigms the morphological analysis of adjective has been done through the proposed system that receives an input in the form of a sentence. The proposed methodology for the morphological analysis of Dogri adjectives using the paradigm approach is structured into the following steps:

4.1 Paradigm Construction

After collecting the data, adjectives are organized into paradigms based on shared morphological patterns. For example, adjectives are grouped according to their inflectional behavior in response to grammatical features like gender (masculine, feminine), number (singular, plural), and case. Each paradigm is analyzed for its base forms (stems) and the specific affixes or modifications applied during inflection. To construct paradigms, a corpus of approximately 15830 unique words has been used. From this sample data, comprehensive lists of all possible suffixes are compiled. Words that share a common set of suffixes are organized into a single paradigm. For instance, words like "□□□□□", "□□□□□□" and "□□□□" all undergo the same inflectional changes, so they are grouped together within the same paradigm. After the construction of adjective paradigms in Dogri, these paradigms can be categorized into four distinct categories based on their inflectional patterns and how adjectives change to agree with nouns in terms of gender, number, and case. These categories help organize and understand the variations in adjective forms within the language: Adjective_F, Adjective_M, Adjective_Const, and Adjective_Case.

4.2 Morphological Analysis of Adjectives: The final step involves a detailed analysis of the paradigms to uncover the underlying morphological rules and patterns. On the input, we execute tokenization and stemming as part of the process. Then we extracted both the root and suffix values from the stemmer, and then, using the root word and suffix as a reference, determined the correct paradigm through a dictionary lookup. The proposed model examines the grammatical characteristics of the input inflected words by considering both the paradigm and suffix value.

4.3 Algorithm

- **Procedure** Adjectives_identification(dataset, tokenization, stemmer, extraction)

```
1.  Variables:
2.      Txfl: Textfile.txt
3.      Snt ← extraction(Txfl)
4.      Morph ← tokenization(Snt)
5.      Morphstm ← stemmer(Morph)
6.      Resultadj : String
7.  Begin:
8.     Snt [j] ← extraction(Txfl)
9.     While Snt[j] ≠ EOF do
10.       Morph[i] ← tokenization(Snt[j])
11.       While i < Morph.len do
12.          Morphstm[k] ← stemmer(Morph[i])
13.          i ← i + 1
14.        k ← k + 1
15.       While k < Morphstm.len
16.        If (dataset.Dictionary(Morphstm[k])) = True then
17.  pgdm[k]=Dictionary.Get_pgdmvalue(Morphstm[k])
18. res[k]=Dictionary.Features(pgdm[k],Suffix() Morph[k]), Morphstm[k])
19.    Display(res[k])
```

```
20.      else
21.          Display("Unknown Word")
22. end:
```

5. Experimentation Results and Analysis

The experimental evaluation for the morphological analysis of Dogri adjectives involves training the system on a dataset comprising 15,830 unique words and testing its performance on a separate set of 1,426 words. During the testing phase, the system's ability to correctly identify and analyze the grammatical features of adjectives, such as gender, number, and case, is assessed using key performance metrics: Precision, Recall, and F1-Score. These metrics provide a comprehensive understanding of the system's accuracy, robustness, and consistency in handling linguistic complexities.

S. no.	No. of words	Recall	Precision	F1-Score
Set-1(Vocative)	85	0.716	0.802	0.757
Set-2(Oblique)	148	0.735	0.826	0.778
Set-3(Direct)	256	0.761	0.813	0.786
Set-4(Singular)	392	0.748	0.806	0.776
Set-5(Plural)	545	0.725	0.798	0.760

The experimental results, which are detailed in the accompanying table, showcase how the model performs when tested on varying dataset sizes. The analysis of the results presented in the table highlights the performance of the system for the morphological analysis of Dogri adjectives across five distinct word sets: Vocative, Oblique, Direct, Singular, and Plural. The system demonstrates strong performance in analyzing Dogri adjectives, with high Precision, Recall, and F1-Scores across grammatical categories. The Direct set achieves the highest F1-Score (0.786), indicating robust handling of this case, while the Oblique set shows the highest Precision (0.826), and reflecting accurate predictions. The Singular and Plural sets exhibit balanced metrics, though the Plural set reveals challenges in identifying complex morphological forms, with slightly lower Recall (0.725). The Vocative set, despite high Precision (0.802), has room for improvement in Recall (0.716). The analysis of experimental results, which includes pattern identification, stemming, paradigm validation, and the identification of grammatical features, enriches our understanding of the language's complexity with which we conclude that the model performance depends on the size of the dictionary, the quantity of the paradigm constructed and mapping of the words to their corresponding paradigms.

6. Conclusion

The morphological analysis of adjectives using a paradigm approach represents a systematic and effective method for understanding the inflectional patterns of adjectives in Dogri language. Through the examination of experimental results and the development of linguistic paradigms helps researchers to gather information about how adjectives adapt to convey gender, number, and case information. The proposed system achieved a balanced performance overall, with high Precision, Recall, and F1-Scores across most sets, particularly excelling in the direct and oblique categories, which showcased the highest F1-Scores. However, there were some irregularities to

the model performance due to the reason that some adjectives may simply exhibit a high degree of inflectional variability, making it challenging to predict their forms based on standard paradigms. These limitations highlight the need for further development and refinement of the system to improve its accuracy and performance.

References

[1] M. Tanaka and H. Yamamoto, "A Corpus Study of Emotive Adjectives and Verbs of the Heian Japanese," 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Kyoto, Japan, 2012, pp. 377-380, doi: 10.1109/SNPD.2012.101.
[2] D. M. Harikrishna and K. S. Rao, "Classification of children stories in hindi using keywords and POS density," 2015 International Conference on Computer, Communication and Control (IC4), Indore, India, 2015, pp. 1-5, doi: 10.1109/IC4.2015.7375666.
[3] V. Goyal and G. S. Lehal, "Hindi Morphological Analyzer and Generator," 2008 First International Conference on Emerging Trends in Engineering and Technology, Nagpur, India, 2008, pp. 1156-1159, doi: 10.1109/ICETET.2008.11.
[4] G. Dhopavkar and M. Kshirsagar, "Syntactic analyzer using morphological process for a given text in natural language for Sense Disambiguation," 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence), Noida, India, 2014, pp. 911-914, doi: 10.1109/CONFLUENCE.2014.6949378.
[5] R. C., D. K., R. Ravindran and K. P. Soman, "Rule Based Reordering and Morphological Processing for English-Malayalam Statistical Machine Translation," 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, Bangalore, India, 2009, pp. 458-460, doi: 10.1109/ACT.2009.118.
[6] D. Chakrabarti and P. Bhattacharyya, "Syntactic alternations of Hindi verbs with reference to the morphological paradigm," Language Engineering Conference, 2002. Proceedings, Hyderabad, India, 2002, pp. 77-84, doi: 10.1109/LEC.2002.1182294.
[7] W. A. Narejo and J. A. Mahar, "Morphology: Sindhi morphological analysis for natural language processing applications," 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, 2016, pp. 27-31, doi: 10.1109/ICECUBE.2016.7495248.
[8] J. Sheth and B. Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language," 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014, pp. 151-154, doi: 10.1109/ICICICT.2014.6781269.
[9] P. Das and A. Das, "Bengali Noun Morphological Analyzer," 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 2013, pp. 1538-1543, doi: 10.1109/ICACCI.2013.6637408.
[10] R. Haque, S. Penkale, J. Jiang and A. Way, "Source-Side Suffix Stripping for Bengali-to-English SMT," 2012 International Conference on Asian Language Processing, Hanoi, Vietnam, 2012, pp. 193-196, doi: 10.1109/IALP.2012.61.