Domain Adaptation in Morphological Analysis

Prathyusha Kuncham, Chandu Khyathi Raghavi, Kovida Nelakuditi, and Dipti Misra Sharma

***Abstract*—Morphological Analysis is a major component in many Natural Language Processing (NLP) applications. The performance of general purpose morphological analyzer (GMA) degrades when used for a particular domain. In this paper we present our effort in developing a domain specific morphological analyzer (DMA) whose architecture is an extension of the existing paradigm based GMA. The method involves identifying domain specific words from a raw text and assigning a paradigm class to them. The proposed method is language independent and has been tested on domain specific Hindi data. The results show 90.60% coverage which is an increase by 6% over GMA and accounts for 25.39% of unanalyzed words.**

***Index Terms*—Morphology, domain adaptation, paradigm table.**

I. Introduction

Morphological analysis is an important step for any linguistically informed NLP application. Morphology is of two types, Inflectional morphology and Derivational morphology.

Inflectional morphology: It is the study of those processes of word formation where various inflectional forms are formed from the existing stems. For example, the English word "cars" is a noun that is inflected for number, specifically to express the plural; the content morpheme "car" is unbound because it could stand alone as a word, while the suffix "-s" is bound because it cannot stand alone as a word. These two morphemes together form the inflected word "cars".

Derivational morphology: It is the study of those processes of the word formation where new words are formed from the existing stems through the addition of morphemes. The meaning of the resultant new word is different from the original word and it often belongs to a different syntactic category. For example, adding "–ful" to "beauty" changes the word from a noun to adjective (beautiful). The form that results from the addition of a derivational morpheme is called a *derived word* or a *derivative*.

Hindi is a morphologically rich language when compared to languages like English. In other words, the morpheme to word ratio is comparatively high in Hindi with respect to English. Also, there is a relatively high chance to find more than one feature in a morpheme in Hindi. For example, 'khAthA' (meaning: eat (in habitual mode)), has two morphemes 'khA' (meaning: eat) which is the root and 'thA', which incorporates the information of gender — male, number — singular and mode — habitual. Thus in this example, we see a single word conveying four morphological features and also a single morpheme conveying three morphological features. The equivalent word for 'khAthA' in English is 'eat'. It conveys information about three morphological features including root — eat, tense — present, person — first/second. As we observe, the morphological features to word ratio for 'khAthA' in Hindi is 4:1 and for 'eat' in English is 3:1.

Borrowing refers to the process of speakers adopting words from a source language into their native language. It is a consequence of cultural contact between two language communities. Borrowing of words can go in both directions between the two languages in contact, but often there is an asymmetry, such that more words go from one side to the other. In this case the source language community has some advantage of power, prestige and/or wealth that makes the objects and ideas it brings desirable and useful to the borrowing language community.

Example 1: English speakers adopted the word *garage* from French, at first with a pronunciation nearer to the French pronunciation than is now usually found.

Example 2: The word 'verandah' from Hindi 'varanda', which probably is from Portuguese 'varanda', originally "long balcony or terrace," of uncertain origin, possibly related to Spanish 'baranda' (meaning: railing), and ultimately from Vulgar Latin barra [1] (meaning: barrier or bar). French 'véranda' is borrowed from English.

Borrowing of words from other languages is widespread in Hindi, so as in many other Indian languages. Particularly in the gadget domain, it is easier to borrow the names of gadgets and their features from English, Latin etc. This poses additional challenge in morphological analysis for a domain with high degree of borrowed words or terms.

Domain adaptation is a common problem in many NLP applications such as Machine Translation, Dialogue systems etc. Further, words native to the language also show a different behavior in a different domain. For example a GMA for Hindi would tag "nihArI" (watch) as verb where as a DMA for recipe domain would tag "nihArI" (a type of dish) as noun. This shows that analyzing domain specific words is very important to capture the correct meaning of a word in a particular domain. The GMA may not identify all the domain specific words as it is trained on the generalized corpus.

In this paper, we present our work regarding the development of a DMA for Hindi. A DMA gives morphological analysis for words in that particular domain. The words in a domain can be broadly divided into two

Manuscript received August 2, 2014; revised April 24, 2015. This work was supported in part by the Language Technologies Research Center of International Institute of Information Technology, Hyderabad.

The authors are with the International Institute of Information Technology, Hyderabad, India (e-mail: prathyusha.k@research.iiit.ac.in, chandukhyathi.raghavi@research.iiit.ac.in, dipti@iiit.ac.in, nelakuditi.kovida@research.iiit.ac.in).

[1] before the word indicates the protoform of the word, which is a hypothetical form of a word, reconstructed from derived words.

categories.

1) Domain specific words which are native to the language.
   Examples: 'malAyi' (cream), in Hindi is specific to recipe domain.
2) Borrowed words from other languages. This is further divided into two types:
   - Borrowed words that take Hindi inflections:
   Examples: 'imeja' (image) which takes 'imejoM' (images).
   - Borrowed words that do not take Hindi inflections:
   Examples: 'grAPiks' (graphics), 'dual' (dual)

Table I shows the distribution of unanalyzed words in above mentioned categories. From the table, we can infer that GMA does not analyze most of the domain specific words.

TABLE I: CATEGORICAL DISTRIBUTION OF WORDS

Number of sentences	1000
Number of tokens	8799
Words unanalyzed by GMA	722
Unanalyzed domain words with Hindi inflections	352
Unanalyzed domain words without Hindi inflections	370
% of unanalyzed words (error %)	8.21%

The rest of the paper is organized as follows. In Section II, we discuss the related work in this field. In Section III, we describe the working of the existing GMA. In Section IV, we propose our approach to develop the DMA. Finally we discuss the evaluation and results of our experiments before we give the conclusion in Section VI.

II. RELATED WORK

A DMA for Hindi has not been developed so far. There are GMAs [1] which uses a rule-based approach that takes both prefixes as well as suffixes into account. They use a corpus and a dictionary to obtain a set of suffix-replacement rules for deriving an inflected word's root form.

Ref. [2] follows a database driven approach which stores the data in normalized form in tables, instead of using files and developed a GUI based morphological analyzer and generator for Windows platform.

A derivational morphological analyzer for Hindi language is presented in [3], which successfully incorporates derivational analysis in the inflectional analyzer and also increases the coverage of the inflectional analysis of the existing inflectional analyzer.

Ref. [4] undertook a comprehensive evaluation of Paradigm Based Approach using the data from the Hindi Tree Bank and presented a new morphological analyzer trained on the Hindi Tree Bank. It is a statistical analyzer that has better coverage and accuracy when compared to existing GMAs in Hindi.

SMA++ i.e., [5] is an improvement over the statistical morph analyzer (SMA) [4]. They modified SMA by adding some rich machine learning features characterized for Indian Languages.

Ref. [6] is also a statistical morphological analyzer which gives state-of-art performance in Hindi.

Attempts at domain adaptation have been done for English language. One such experiment [7] was performed by Hal Daum'e III, at the School of Computing, University of Utah. They use a fully supervised learning approach which is based on augmentation of features that are extracted for each word.

III. WORKING OF EXISTING GMA

A paradigm based GMA has been developed at IIIT Hyderabad [8]. It uses both paradigms and a root dictionary to provide inflectional analysis of a word. The analyzer handles inflected words using the paradigm tables. Table II shows the word-forms of ladakA (boy).

Table III is the paradigm table for 'laDakA' (boy). The paradigm table contains add-delete rules which are used to compute the inflections of the word. For example by looking at Table III we can say that oblique, singular case of 'laDakA' (boy) can be formed by using the rule e/A i.e. 'laDakA' -A +e = 'laDake'. Using the inverse of the add-delete rules, we can get a root word which is used for finding morphological analysis of the word. All the words that behave alike i.e. follow the same add-delete rules to form all their respective word-forms belong to the same paradigm. 'kapaDA' (cloth), 'gODA' (horse) etc., follow the same paradigm 'laDakA'(boy). Paradigms are designed in such a way that each root word belongs to only one paradigm.

TABLE II: WORD-FORMS OF LADAKA (BOY)

Number/Case	Singular	Plural
Direct	laDakA	laDake
Oblique	laDake	laDakoM

TABLE III: PARADIGM TABLE OF LADAKA (BOY)

Number/Case	Singular	Plural
Direct	-/-	e/A
Oblique	e/A	oM/A

Every line in the root dictionary contains a root, its paradigm and other grammatical information of the root. If a word is present in the root dictionary, the analyzer handles all the inflections pertaining to that word using the corresponding paradigm. For example: 'le' (take) is a root word present in the dictionary. The other word-forms of 'le' (take) are 'lethA' (takes), 'liyA' (took), 'lenA' (to take) etc., are also handled by the analyzer without requiring an explicit entry in the dictionary. So a word can be analyzed only if its root word is present in the root dictionary.

The GMA's output is a feature structure which contains features such as root, lexical category, gender, number, person, case, vibhakti and tense-aspect-modality (TAM) of the input word.

Example:

Input: 'laDakoM' (boys)

Output: <fs af='laDakA, *n*, *m*, *pl*, 3, *o*,,'>.

Here the root is 'laDakA' (boy), '*n*' is the category (Noun), gender is masculine ('*m*'), person is '3', number is plural ('*pl*') and case is oblique ('*o*').

The above discussed GMA works as a platform for the DMA. Our tool not only gives morphological analysis of the words whose root forms are present in the root dictionary but also analyzes certain borrowed words whose root forms are not present in the dictionary.

IV. Our Approach

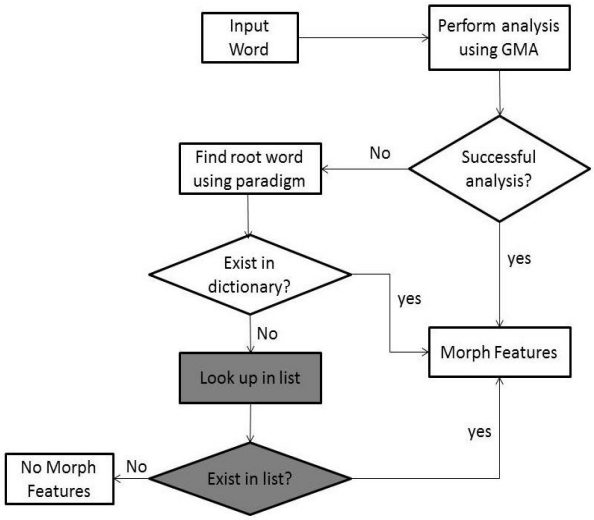


Fig. 1. Internal working of DMA.

Our approach can be mainly divided into three steps.

1) Identifying domain words.
2) Analyzing words with Hindi inflections.
3) Analyzing words that do not take Hindi inflections.

Fig. 1 shows a flowchart of internal working of DMA. In the flow chart, dark coloured shapes show the analysis of words that do not take Hindi inflections.

A. Identifying Domain Words

1) We collected gadget data of about 10,000 sentences from news-paper articles and tokenized them.
2) The words/tokens collected are given as input to the GMA for Hindi. From the output, the words which are not identified i.e. marked as unknown ('unk') by GMA are collected.
3) From such unknown words, only the words that belong to gadget domain are considered. Examples of domain words: 'gejetoM' (gadgets), 'lAptop' (laptop) etc.

Having identified the domain data, we process them based on following steps.

B. Analyzing Words with Hindi Inflections

1) If all the word-forms of an input word existing in the language can be generated by the add-delete rules of a paradigm, then automatically that paradigm is assigned to the root word (as discussed in Section III, root word can be generated from inverse add-delete rules). An entry with the root word, its paradigm and other grammatical features is made in the dictionary.

TABLE IV: Generation of Word-Forms of 'cAta' (Chat) from 'ghara' (House) Paradigm

Type	Word forms of the paradigm	Add-delete rules	Word forms of 'cAta'
Direct Singular	ghara	-/-	cAta
Direct Plural	ghara	-/-	cAta
Oblique Singular	ghara	-/-	cAta
Oblique Plural	gharoM	oM/a	cAtoM

For example, consider an input word 'cAta'(chat). The existing forms of 'cAta' in Hindi are 'cAta' and 'cAtoM'(chats). From the Table IV, these can be generated by the add delete rules of 'ghara'(house) paradigm. Hence 'cAta' is assigned to 'ghara' paradigm.

2) In cases where a few or none of the word-forms of a particular word occur in the corpus, one of the existing paradigms was manually assigned and a corresponding entry was made in the dictionary.

C. Analyzing Words That Do Not Take Hindi Inflections

The words which take Hindi inflections are analyzed in Section IV, part 2 and we are left with borrowed words that do not take Hindi inflections. These words are stored in a list along with their possible feature structures.

Example: A word like 'gejatasa' (gadgets) which is a borrowed word that doesn't take Hindi inflections is stored in a list with its feature structures.

V. Experiments and Results

We tested on 857 sentences of gadget data taken from newspaper articles. The total tokens/words in these sentences are 10596.

Table V shows the percentage coverage of words i.e. percentage of words analyzed by GMA and DMA.

TABLE V: Comparison of Results between GMA and DMA

Type of morph analyzer	Total unknown words	Unique unknown words	% of coverage
GMA	1627	571	84.64
DMA	995	426	90.60

From the Table V, we can see that there are a total 1627 words that are not analyzed by GMA and only 995 words are remained unanalyzed by DMA, i.e., 632 more words are analyzed by DMA than GMA. Out of these the unique unknown words are 571 and 426 respectively.

We observe that there is approximately 6% increase in the coverage of words using the DMA over GMA. We also observe that out of 571 unique unknown words that were left unanalyzed by GMA, 145 (571-426) were analyzed by DMA which amounts to 25.39% improvement.

Based on the results from Table V, among the 426 unique words that were not identified by DMA, around 100 words (23.47%) were named entities. So if a named entity recognizer is incorporated along with DMA, it would yield better results.

A borrowed word can be written in different spellings, which tampers the performance of the system. For example, 'AIkoyn', 'Aikon', 'AIkOn' all refer to the word 'icon'. So the use of a normalizer would be very beneficial in this case because it maps different spellings of a word to a single normalized spelling.

VI. Conclusion

We developed a plug-in over GMA to analyze domain specific words. It expands the coverage of GMA. We also observed that among the words that have been covered by

DMA, there were certain domain words that are extremely essential to convey domain information which were left unanalyzed by GMA. This result is useful not only in acquiring morphological analysis but also in improving the accuracy of all other modules which require morphological information.

The DMA can be readily adapted to other languages and other domains that have to deal with morphological analysis. In general, only language specific parts have to be replaced for this purpose. We are continuing our efforts to fully automate the Domain Specific Morphological Analyzer incorporating Machine Learning techniques. As discussed in Section V, additional tools such as Named Entity Recognizer and normalizer can also be added to improve the flexibility and coverage of the domain model. We consider this as our future work.

ACKNOWLEDGMENT

The research conducted in this project has been supported by Language Technology Research Center (LTRC) at International Institute of Information Technology, Hyderabad (IIITH), India. We also thank them for making available morph developed at IIIT freely downloadable to NLP researchers.

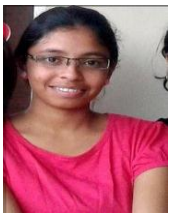
We would like to thank Mehnaaz Mohiuddin for her contribution to the work and Himani Chaudhry for her valuable advices throughout the writing of the paper. A special mention of thanks to Himanshu Sharma who has helped with the presentation of the paper.

REFERENCES

[1] N. Aswani and R. J. Gaizauskas, "Developing morphological analysers for South Asian languages: Experimenting with the Hindi and Gujarati languages," *LREC*, 2010.
[2] V. Goyal and G. S. Lehal, "Hindi morphological analyzer and generator," in *Proc. First International Conference on Emerging Trends in Engineering and Technology*, 2008.
[3] N. Kanuparthi, A. Inumella, and D. M. Sharma, "Hindi derivational morphological analyzer," in *Proc. the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, 2012.
[4] D. K. Malladi and P. Mannem, "Statistical morphological analyzer for hindi," in *Proc. 6th International Joint Conference on Natural Language Processing*, 2013.
[5] S. Srirampur, R. Chandibhamar, and R. Mamidi, "Statistical morph analyzer (SMA++) for Indian languages," *COLING*, vol. 103, 2014.
[6] D. K. Malladi and P. Mannem, "Context based statistical morphological analyzer and its effect on Hindi dependency parsing," in *Proc. Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, vol. 12, 2013.
[7] H. Daum é (July 2009). Frustratingly easy domain adaptation. [Online]. Available: http://arxiv.org/abs/0907.1815
[8] A. Bharati, V. Chaitanya, and R. Sangal, *Natural Language Processing: A Paninian Perspective*, New Delhi: Prentice-Hall of India, 1995.

Prathyusha Kuncham was born on December 23, 1992 in Hyderabad, Telangana State, India. She completed her B.Tech degree in computer science and engineering from International Institute of Information Technology (IIIT) and she is currently pursuing her MS degree by research in computational linguistics in International Institute of Information Technology, Hyderabad, India.

Chandu Khyathi Raghavi was born in Andhra Pradesh State, India. She is currently pursuing her fourth year of under graduation in an integrated five year dual degree program, B.Tech degree in computer science and MS degree by research in computational linguistics, in the Language Technologies Research Centre of International Institute of Information Technology, Hyderabad.



Kovida Nelakuditi was born on December 29, 1993 in Guntur, Andhra Pradesh State, India. She is in the fourth year of an integrated program, B.Tech degree in computer science and MS degree by research in computational linguistics in International Institute of Information Technology, Hyderabad, India.

Dipti Misra Sharma is the head of the Language Technologies Research Center and a professor at International Institute of Information Technology, Hyderabad (IIIT-H), India. Dr. Sharma did her post-graduation and Ph.D. degree in linguistics from University of Delhi. After spending a year in Germany working on a project on 'Orality in written literatures', she came back and first joined Osmania University and then moved to University of Hyderabad (UoH) as a UGC research associate. Thereafter, she worked as a faculty at UoH teaching various courses in applied linguistics.