

Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages

Akshar Bharati,
Rajeev Sangal, S.M. Bendre, Pavan Kumar, Aishwarya
Language Technologies Research Center
International Institute of Information Technology, Hyderabad
{sangal,bendre}@iiit.net, {pavan,aishwarya}@gdit.iiit.net

Abstract

This paper presents an algorithm for unsupervised learning of morphological analysis and generation of inflectionally rich languages like Hindi, given a low coverage morph and a corpus of raw text. It assumes no particular theoretical model of morph, but can work with any morph that defines classes of stem that behave similarly. The morph learning algorithm uses the concept of 'observable paradigm'. The results of the algorithm are encouraging with the coverage of a primitive morph going up from 32% to about 63% and that of an advanced morph going up from 96% to about 97%.

1. INTRODUCTION

This paper presents a morph learning algorithm which primarily uses the frequency of occurrences of word forms in a raw corpus. It depends on the empirically verified assumption that the proportion of frequencies of word forms for different feature structures are comparable across stems. A paradigm specifies a common morphological process for a set of given stems. We introduce the concept of 'observable paradigm' by forming equivalence classes of feature-structures

which are not distinguishable.

In particular, when an unknown word is encountered, the following steps are followed:

1. A guessing component proposes (stem, paradigm) pairs.
2. For each pair, a set of word forms is generated and the corresponding feature-structure set is partitioned into equivalent classes based on the concept of observable paradigm.
3. Frequencies of the word forms for each equivalence class in the partition are obtained and a suitable vector created.
4. Finally, such vectors are compared for each guess with respect to pre-compiled reference vectors to select the most likely (stem, paradigm) pair.

The method is similar to Yarowsky and Wicentowski (2000), but it takes care of the cases not handled by them pertaining to spelling variation and overlapping word forms for different feature structures. The notion of observable paradigms is introduced precisely for this purpose. Rather than deal with individual word forms for a stem, we deal with classes as obtained by observable paradigms. The recent work of Goldsmith (2001) is interesting in that the no-

tion of minimum description length (of morphological rules etc) is used to introduce machine learning of the morphology of the language. However, because additional knowledge source besides a raw corpus is not assumed, (not even a primitive morph), learning is confined to signatures, without trying to carry out further inference. Another recent contribution is by Oflazer et al (2001), which tries to base itself on human elicitation and machine learning.

The results of the algorithm are encouraging with the coverage of the primitive morph going up from 32% to about 63% and that of the advanced morph going up from 96% to about 97%.

The present framework based on observable paradigms places other work done in this area on a sound theoretical foundation.

2. PARADIGM BASED MORPHOLOGY

This paper assumes no specific theoretical model behind the given morphological package; as a result it can work across any underlying model of the morph. All it assumes is that the morph works using paradigms, where

1. A paradigm defines all the word forms of a given stem and also provides a feature structure with every word form.
2. For every stem in the lexicon, the paradigm it follows is given.

In other words, a paradigm P_i specifies the word forms of a stem, and for each paradigm a list of stems T_i is known which obey the paradigm. The paradigm is used both for word analysis as well as word generation. The paradigm could be implemented in terms of rules, constraints, or tables; the proposed morph learning algorithm is independent of the specific model adopted for the paradigm.

2.1. GIVEN MORPHOLOGICAL ANALYZER

The morph package for Hindi, on which the proposed method was tested, works with paradigms based on add-delete strings. All the word forms of a stem together with their corresponding feature structures are given. The given morph package extracts add and delete strings from the above word forms, and associates them with the corresponding feature structures. The add-delete strings represent suffix information. This extracted information is a paradigm (Bharati et al, 1995). Figure 1 gives the word forms of stem laDakA(boy) along with add-delete strings in the bracket.

<i>Number</i>	<i>Case</i>	
	<i>direct</i>	<i>oblique</i>
Singular	laDakA(0, ϕ)	laDake(1,e)
Plural	laDake(1,e)	laDakoM(1,oM)

Figure 1: Word forms with add-delete string for the stem laDakA

All the distinct paradigms in the language need to be defined (by giving example word forms for a stem). Besides this, a list of pairs consisting of a stem together with the paradigm it follows, is all that is needed to perform morphological analysis or generation. For example, the stems kapaDA, ghoDA etc. follow the paradigm 'laDakA'.

That this model works well for inflectional morphology with moderately rich inflectional forms is discussed in Bharati et al (1995; Chapter 3). As described above, the given Hindi morph package partitions the set of stems into (morphological) classes, and associates a paradigm with each class.

2.2. OBSERVABLE PARADIGM

A paradigm P has a paradigmatic stem S and a set of pairs, each pair consisting of

a feature structure fs and non-empty word set ws for the paradigmatic stem $(fs, ws) = (fs, \{w_1, w_2, \dots\})$. Multiple forms in $ws = \{w_1, w_2, \dots\}$ usually represent spelling variants of a word form. Furthermore, each feature structure (fs_i) in the set of pairs in a paradigm $\{(fs_1, ws_1), \dots, (fs_m, ws_m)\}$ is distinct. The word forms in word sets ws_i and ws_j , across the different feature-structures might be overlapping in the sense, $ws_i \cap ws_j \neq \phi$, that is, there is at least one word form, say w , which occurs in both sets. Whenever two word-sets are overlapping it is not possible to observe the corresponding feature-structures directly. In some sense, it is not possible to distinguish among the occurrences of the word forms corresponding to the feature structures fs_i and fs_j in all cases. Frequency of w obtained from the corpus is a composite frequency of two feature structures. For instance, in English, some of the forms of travel can be distinct (travel and travels), whereas some others (participle and past) can be the same (traveled), as shown in Figure 2, where feature structure of first row is {TAM = present tense, number = pl} and the feature structure of other entries can be similarly specified.

Word Sets	
{take}	{travel}
{takes}	{travels}
{took}	{travelled, traveled}
{taken}	{travelled, traveled}

Figure 2: Example word sets

We form equivalence classes of feature structures for each paradigm as follows: If any two feature-structures fs_i and fs_j have overlapping word sets, ($ws_i \cap ws_j \neq \phi$) then they are in the same equivalence class

$\{fs_i, fs_j\}$. Thus, if fs_i and fs_j overlap in word w , and fs_j and fs_k overlap in word v , all three are in the same equivalence class $\{fs_i, fs_j, fs_k\}$.

If there is an equivalence class consisting of fs_i and fs_j , we have $(\{fs_i, fs_j\}, ws_i \cup ws_j)$. Suppose at the end of the above procedure we have k such equivalence classes ($k \geq 1$). The equivalence classes stand for “distinguishable” and hence “observable” feature structures. Let us call it observable paradigm or OP which represents $\{(fs_1, ws_1), \dots, (fs_k, ws_k)\}$.

Suppose there are d paradigms P_i , $i = 1, \dots, d$ in the morphological analyzer. For each paradigm P_i we have observable set OP_i of k_i pairs say. For the sake of convenience, we assume that for every i , the set OP_i is ordered by some lexicographic scheme on the feature-structures.

In addition to OP_i , corresponding to each paradigm P_i the existing morph also has a list of stems T_i , $i = 1, \dots, d$, which obeys the paradigm, as mentioned earlier.

We assume that the morph package (which includes both the analyzer and the generator) provides the following three functionality:

1. Analysis: Given a word W , the morph analyzer gives an answer consisting of a stem and a feature structure. (In case of ambiguity, there may be several answers, i.e., pairs of stem and feature structure.)
2. Guess: If the morph analyzer is not able to analyze a given word, the morph “guessing” package returns a set of pairs each pair consisting of a stem and a paradigm. (The guessing is based on suffixes, though the details are not discussed in this paper.)
3. Generation: Given a stem S and a paradigm P , it can generate all its

forms together with the associated feature structure with each form.

Here is an example showing the feature structures and word forms for KA(eat) verb in Hindi for TAM equal to future. The feature structure of the first entry is {TAM=future, gender=f, number=p, person=3} and the feature structures of the other entries can be similarly specified.

No.	Word forms
1	{KAezgI/KAyeMgI}
2	{KAogI}
3	{KAiegA/KAiyegA}
4	{KAezgI/KAyeMgI}
5	{KAegI/KAyegI}
6	{KAegI/KAyegI}
7	{KAiegA/KAiyegA}
8	{KAUzgI}
9	{KAezge/KAyeMge}
10	{KAoge}
11	{KAiegA/KAiyegA}
12	{KAezge/KAyeMge}
13	{KAegA/KAyegA}
14	{KAegA/KAyegA}
15	{KAiegA/KAiyegA}
16	{KAUzgA }

Equivalence classes are formed for feature structures producing the observable paradigm. For instance, in the figure above, (1,4) is an equivalence class for the word forms {KAezgI/KAyeMgI} and the equivalence classes of feature structures are {(1,4), 2, (3,7,11,15), (5,6), 8, (9,12), 10, (13,14), 16}.

3. CORPUS BASED LEARNING

The system is trained on an available corpus as follows: For every stem $t \in T_i$, its distinct word forms corresponding to the paradigm P_i are generated using OP_i . The

frequency of each of these word forms, generated for the specified stem and the specific paradigm, is collected from the corpus. These frequencies are summed over words in each word set in OP_i , giving a vector of frequencies, which we denote by C_{t,P_i} . We compute the sum of elements of this vector, say sum of frequencies denoted by F_i , and those stems with F_i less than a threshold value U are deleted from T_i , where U is determined based on the corpus size. Henceforth, by T_i , we denote this reduced list of stems.

For each C_{t,P_i} , a ratio vector D_{t,P_i} is constructed where the entries in D_{t,P_i} are the ratios of frequencies to the sum of frequencies F_i of all word forms of that t . The average of these vectors over all $t \in T_i$ is denoted by AD_{P_i} . The vector AD_{P_i} is a representative ratio vector of the paradigm P_i for every i and we make the following assumption:

The vector $D_{t,P}$ corresponding to the stem t following a paradigm P will be similar to the average vector AD_P and distinct from other average vectors AD_Q ($P \neq Q$).

This assumption was empirically verified with certain verb sets for the test language Hindi. Based on this assumption, we propose the algorithms in the next section.

4. PROPOSED METHOD

Suppose the morphological analyzer comes across an inflectional form which it can not recognize. If a possible stem of the inflectional form and a paradigm can be guessed, the morphological generator will be able to generate the different word forms of the stem. The information available from the corpus is the corpus frequency of the various possible word forms generated from the suggested/guessed stem. The proposed algorithm makes use of the corpus frequencies to

decide on the most likely stem and the corresponding paradigm. Another important factor in choosing the correct inflectional form is the length of the suffix compared to the proposed stem (Goldsmith, 2001) and this fact is taken into consideration while arriving at the decision. In addition, if a specific root occurs with fair amount of frequency in the corpus, then most of its possible inflectional forms specified by the correct paradigm it follows are likely to be present in the corpus. Hence the proportion of number of inflectional forms present in the corpus out of the total number of possible inflectional forms of the specific paradigm under consideration is also taken into account.

In particular, the algorithm carries out the following steps to collect the information from the corpus and arrive at a decision regarding the stem and paradigm of a word not recognized by the morph.

Algorithm

- Given an inflectional form not recognized by the morph (called input word W henceforth) use the morph guessing package to return the set of s possible (stem, paradigm) pairs $(S_1, P_1), \dots, (S_s, P_s)$ which could have generated W .
- For each guessed stem S_i generate all possible distinct inflectional forms k_i based on observable paradigm OP_i corresponding to P_i using the morphological generator, $i = 1, \dots, s$.
- Construct the corresponding ratio vectors D_{S_i, P_i} by collecting the corpus frequencies.
- The vectors D_{S_i, P_i} are compared with the ratio vector AD_{P_i} of the particular paradigm P_i . The most likely pair (S_i, P_i) is chosen based on

$$\operatorname{argmax}_{i=1, \dots, s} (l_i^3 b_i^2) / \langle D_{S_i, P_i}, AD_{P_i} \rangle$$

where, l_i = length of the suffix of W compared to stem S_i . b_i = proportion of inflectional forms out of k_i which have non-zero frequencies in the corpus, and $\langle a, b \rangle$ = the appropriately chosen distance measure between the two ratio vectors a and b .

The two distance measures chosen were i) $F_i^{-2} \times$ Kullback-Liebler distance (KL) and ii) Euclidean distance (Eucl), where Kullback-Liebler distance $(X, Y) = \sum_j x_j \log(\frac{x_j}{y_j})$, and Euclidean distance $(X, Y) = \sum_j (x_j - y_j)^2$.

4.1 SINGLE INFLECTIONAL FORM

In spite of rich inflectional variation in the language, certain words do not change their form. In other words, paradigms have single inflectional form (SIF) across all feature structures (that is, $k_j = 1$). This is particularly true of many noun forms in Hindi, such as vAyu (air), on which the algorithm was tested. The technique used above fails in such cases since both D_{S_j, P_j} and AD_{P_j} are unit vectors of length 1.

To overcome this problem, the word W which has at least one suggested pair (S_j, P_j) with $k_j = 1$ is treated differently. If corresponding to other suggested pairs, the number of inflectional forms of W present in the corpus is larger than one, all pairs with $k_j = 1$ are ignored and the proposed algorithm is carried out. Else, W is considered to have SIF.

4.2 NON-CLASSIFICATION

In certain situations, the argmax may occur for more than one pair (S_j, P_j) , resulting in a clash. Such a word is considered 'not classified'. However, whenever it is possible to select the most likely pair based on absolute corpus frequencies, that particular pair is selected as the correct one. (In such a sit-

uation, the user response could be elicited to resolve the clash.)

For instance, in case of Hindi verbs with TAM = future, the clash often occurred between the paradigms uTa(get up) and le(take) and uTa was chosen based on its high likelihood.

5. PERFORMANCE EVALUATION

To evaluate the performance of the proposed method, the algorithm was tested on 8 Hindi texts of total word count 9804, where the smallest text was of 772 word counts and the largest was of 1662. For Hindi, the latest morphological analyzer is claimed to have total coverage of about 85%. An earlier version of this analyzer is also available which has the total coverage of about 30% only. Both the available morphological analyzers were used to evaluate the performance of the proposed method.

To begin with, both Old and New morph were run on the testing documents. The performance statistics of the two morphs are presented in first three rows of Table 5.1.

Table 5.1: Morph Coverage in no. of words

Morph→	Old	New
Total no.	9764	9764
No. analyzed	1133	8692
	(11.60%)	(89.02%)
No. not analyzed	8631	1072
No. relevant	3514	284
Relevant Coverage	32.48%	96.67%

The words which were not analyzed by the morphs were candidates for the performance evaluation of the proposed method. While applying the method, learning paradigms for the adjectives was not tried, though it is straightforward. Also, in Hindi, the iden-

tification of proper nouns needs to be handled separately. The method at present has not been applied on adjectives and proper nouns. Hence the words falling in these types were not considered for analysis and the last but one row of Table 5.1 gives the number of words considered for analysis, namely the relevant words. Relevant Coverage given in the last row of Table 5.1 is defined to be = (Total no words classified by the morph)/(No of words classified by morph + No of relevant words).

The algorithm was carried out on the relevant words using both the distances. Tables 5.2 and 5.3 below present the performance evaluation of the method.

Table 5.2: Improved Old Morph Coverage

Distance→	KL	Eucl
No. classified	1600	1455
Precision	45.81 %	41.62%
Recall	20.00%	18.08%
Improved Coverage	63.44%	60.65%

Table 5.3: Improved New Morph Coverage

Distance→	KL	Eucl
No. classified	113	104
Precision	40.07%	37.13%
Recall	11.63%	11.34%
Improved Coverage	97.97%	97.88%

where we define Improved Coverage = (Total no of words classified by the morph and algorithm)/(No of words classified by morph + No of relevant words).

The figures presented are the average percentages over 8 texts. There is significant improvement in coverage of Old morph. Also performance based on Kullback-Liebler distance is better than that based on Euclidean distance for all 8 texts considered.

6. DISCUSSION AND CONCLUSIONS

We present an unsupervised improvement of an existing morphological analyzer and generator in this paper. The method assumes that the morphological package makes use of paradigms, where the paradigm does not assume any particular theoretical model of morphology. The package should also be able to guess the stem-paradigm pairs, given an unknown word. By considering equivalence classes of feature structures, the method can handle spelling variations and overlaps of word forms in a language. However this coverage is limited to the variations covered by the existing paradigms in the given morphological analyzer. The method does not propose any new paradigms. The method depends only on the frequencies of word forms in a raw corpus and does not require any input from the user regarding linguistic rules or tagging of the corpus. As a result, the performance of the method largely depends on the corpus size.

The method was tested on the existing morphological analyzer of Hindi along with a raw corpus of size 13 million. It is observed that the method works well even in situations where the list of stems which follow a particular paradigm is small. Hence, the proposed method can be used for a fast improvement of a morphological package which might be at a preliminary stage of development. It can also be used to correct the errors of wrong paradigm labels for stems in an existing morphological package. At present, it fails to classify or label a small percentage of words considered and on eliciting user response for such words, a minimally supervised learning algorithm can be implemented to improve the morph.

The choice of word types was restricted for this study and the word types such as adjectives

and proper nouns were not considered.

Also, for Hindi, the identification of proper-nouns needs to be handled separately. If proper-nouns can be marked using a suitable tool, one can attempt to handle their morphology, which has not been worked out at present.

Few issues which came up out of this study are:

1. Sparse data problem : Since the learning and analysis is entirely based on frequencies, the roots with low frequencies have high chances of getting misclassified. About 10% words of the test data were of this type. However, their contribution to the coverage of the morphological analyzer on an arbitrary text is expected to be small because of low frequency of such words. As part of the future work, it would be interesting to find the relation between the size of the corpus needed (Bharati et al, 1998) and the improvement in coverage.
2. Rare word forms : Certain words have reasonably high frequency of occurrence, but other word forms for the same stem are extremely rare, leading to lack of evidence and mis-classification. For instance, word such as vittAMantri (finance minister) has a rarely occurring correct plural oblique vittAmantriM. The fact that there is only one finance minister in a country seems to contribute to this. About 40% of the misclassified words fall in this category. Although this again does not affect the coverage of the resulting morph substantially, it is pointed out as an issue. There are some deep relations between the world (as it is) and the language (which models or describes the world). This is different from the usual sparse-data problem; a substantial increase in

- the corpus would help only marginally.
3. SIF : Some word forms of a stem for which only one of the forms is found in the given corpus (because of reasons (1) or (2) above) get assigned to SIF paradigm. This error does not contribute to 'mis-classification' of other word forms of this stem because there are no other word forms in SIF. At worst, it would fail to handle the other word forms of the stem and they would remain unclassified. Since the frequency of such word forms is low, their contribution to coverage is small.
 4. Proper nouns : The proper nouns typically behave like (2) above. However, in Hindi, some of the proper nouns are also common nouns, and as common nouns their stems have other forms with good frequency; for instance latA (creeper).

Performance analysis of the present method is being extended to include adjectives, adverbs etc for Hindi. As part of future work, we plan to apply this method to morphological analyzers of other Indian languages. It would be interesting to study the improvement in coverage obtained for languages with different degree of inflectionality. Even more interesting would be to come up with a method of estimating corpus size needed to achieve a given level of coverage of morph for a language. Finally it would be worthwhile to use more than one criterion (namely, frequency of words) in machine learning of morphology. The criteria could be: identification of proper nouns using appropriate tools, values of features such as number, person etc using parser and so on.

7. REFERENCES

1. Bharati, Akshar, Vineet Chaitanya and Rajeev Sangal. (1995). *Natural Lan-*

guage Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi.

2. Bharati, Akshar, Rajeev Sangal and S.M. Bendre (1998). Some Observations on Corpora of Some Indian Languages. *Knowledge Based Computing Systems*, Tata McGraw-Hill.
3. Goldsmith, John. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2), pp 153-198.
4. Oflazer, Kemal, Sergei Nirenburg, and Marjorie McShane, (2001). Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics*, 27(1), pp 60-85.
5. Yarowsky, D. and Wicentowski, R. (2000). Minimally Supervised Morphological Analysis by Multimodal Alignment. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL'00)*, pp 207-216, Hong Kong.