Classification of Words Based on Affix Evidence

Utpal Sharma∗ Jugal Kalita† Rajib Das‡

Abstract

Category of a word, as may be indicated in a lexicon, is a useful piece of information for most linguistic exercises. Hence determination of categories of words in a natural language is an important task. There has been much work in Part-of-Speech tagging of words in texts, where the tagset is usually pre-determined and provided to the system. In this work we attempt to determine the categories of the root words of a language by considering the usage of suffixes in a corpus of a (highly) inflectional language. This information can be used in a lexicon. Instead of using a pre-defined list of word categories, we first identify the underlying word categories by using a set-theoretic representation of the information on suffix usage available from the input corpus.

1 Introduction

Classification of words of a natural language into different linguistic categories is an important task in most linguistic exercises. Though such classification may not be explicitly done by a human user of a language, still there is no doubt that in a formal linguistic analysis, such as syntax or semantic analysis ([1], [17]), the knowledge of the category of each word is important. This knowledge is useful in part-of-speech (POS) tagging in processing a natural language text. In traditional linguistic exercises the categories of the words are determined manually using different sources of information from existing catalogs to context of usage of each word. In computational linguistics the task is not too simple because often the required catalogs are not available, and the mechanism to analyse the context of the usage of the words is not clear. In this paper we present a computational method that uses evidence of affix usage for identifying the underlying classes of words in a language and then classifying words. The classes identified are not exactly POS classes (eg., [12]) used for tagging words in a text, rather they can be used as “category” attribute of the words in a lexicon. Though we have carried out our experiments for Assamese, but we strongly feel that the method is general and applicable for most inflectional languages.

∗Department of Computer Science, Tezpur University, Tezpur - 784028, Assam, India, utpal@tezu.ernet.in
†Department of Computer Science, University of Colorado, Colorado Springs, CO 80933, kalita@pikespeak.uccs.edu
‡Department of Computer Science, Tezpur University, Tezpur - 784028, Assam, India, rkd@tezu.ernet.in
2 Classification of Words

Linguistic categories of words, such as noun, verb, adjective, etc. reflects the meaning of the words, although to a limited extent. For example, if a word is categorized as a noun, then it implies that the word denotes an object, either physical or abstract. (We would like to think that the entire meaning of words can be viewed as successive classification, though classification beyond a point is not explicitly done. For example, the word pen may be classified as a noun, physical, usage - writing, dimension - 10-15 cm long, 1 cm diameter, etc.) In a way this means that classification of words requires the knowledge of the meanings of the words. Many computational methods take the list of word categories or parts-of-speech, in a language, and the criteria in some form to classify the words into these categories, as input (e.g., [5], [4]). But is it possible to classify words when no prior information about either the meanings of the words, or the underlying categories of words, is available? This is often a problem in computational linguistics.

The solution to the above computational problem probably lies in the analysis of the structure of the words and their usage in sentences or phrases. Either or both of these display some patterns that can help in categorizing the words to certain extent, though the final meaning may not be inferred from the structure of the word or sentence alone. Various approaches have been taken by researchers that make use of such observations as well as related statistics for word classification. (eg. [5], [6], [10], [4], [16], [15], [13], [14], [19]). These approaches generally adopt one of the two broad approaches - rule-based and stochastic. Further, they either use unsupervised or supervised training. Most such work, however, target part-of-speech identification, rather than a general context-independent categorization of the words.

In this work we discuss identification of underlying categories for root forms of words in a language and then the classification of words by considering the morphology, and more specifically, the use of suffixes in words. We consider the words without considering the context (neighbouring words), but we consider the various inflected forms of words in a corpus together, to figure out the category of the base word. In a way, the inflected words can be seen as a sequence of the morphemes, and thus an n-gram. But in this method we do not consider each such n-gram to guess the category of the inflected word or the base morpheme. Rather we consider multiple n-grams where the base is same and try to guess the category of the base. And in the process we first figure out the list of categories of words in the language from the same evidence. The categorization of the root words is expected to be useful in a lexicon being built for the language.

3 Classification of words by suffix evidence

Each suffix in a language affects a base word making it suitable to play a specific type of role in a sentence. For example, in the English word play-ed, the suffix -ed makes the base word play suitable to represent an action in the past. It is also understandable that such an effect by a suffix can be expected for words of a certain type of meaning only. In the above example, the suffix -ed applies to words representing action, i.e., verbs. There can be exceptions to this observation. For example, the suffix -s in English applies to verbs and nouns with different roles. By considering the application of suffixes to words carefully, we can guess the categories of base words. Thus, in this discussion our
notion of linguistic category will be such that – two words will be in the same category if the set of suffixes that they can take is same. In an highly inflectional language like Assamese, the word categories identified by the above criteria can be more precise than the usual linguistic categories, such as noun, verb, etc. That is, say, within nouns we may be able to find subclasses. This is often significant - for instance, in Assamese different determiners are allowed with different categories of nouns. The idea of such categorization of words is considered in [7] for the purpose of morpho-syntactic parsing. However, input and output of the system described there is different from those of our method.

Identifying patterns of affixation in words in a corpus, termed as signatures, has been described in [8] where words are grouped on the basis of such patterns invoking the principles of Minimum Description Length (MDL) framework. But since each individual word in a corpus do not always occur with all affixes valid for it, hence such groupings based on direct evidence from a corpus may not hold beyond that corpus. What is required is some further analysis of the affixes. Some simple possible approaches for identifying linguistic categories of words in a corpus by considering suffix evidence have been mentioned in [18]. However, it turns out that the simple ideas need to be strengthened because of two main reasons, namely, several suffixes may apply to words of the same category, and, same suffix may apply to words of different categories. Apart from this, evidence provided by individual words in a corpus, are often sparse. This situation is depicted pictorially in Figure 1 for a hypothetical language L.

![Figure 1: Suffixes and linguistic categories of words for language L](image)

In Figure 1 the letters denote different suffixes and the numbered ellipses enclosing the letters denote linguistic categories of words that occur with those suffixes. The suffixes c, d, e, f, g do not clearly imply any distinct category for the words that take these suffixes, and the suffixes h, i do not indicate different categories. On the other hand, the suffixes a, b, c, h, i unambiguously imply the categories 1, 2, 3, 4, 4 respectively.

In a computational linguistic exercise it is often the case that the input to the system is only a text corpus, and possibly, a break-up of words into base and suffix wherever applicable. Definitions of the word categories in terms of suffixes are not given to the system. The problem then is to identify the underlying linguistic categories of words (1, 2, 3 and 4 in Figure 1) in the language and classify the words into these categories.
4 Pivot Suffixes and Word Classification

In a text corpus with the break-up of words into base and suffix indicated wherever applicable, we shall encounter cases of the same word appearing with different suffixes. However, each word will occur with only a subset of all the possible suffixes of its linguistic category. For example, in a text corpus of language $L$ of Figure 1, we may find different words occurring with the sets of suffixes - $\{a, e\}$, $\{b, f\}$, $\{a, e, d\}$, $\{d, c\}$, $\{a, f\}$, etc. respectively. In other words, we have a set of words, $W = \{w_1, w_2, ..., w_n\}$, and for each word the list of associated suffixes, i.e., suffixes that apply to it.

Let us define co-occurrence of suffixes as follows - a suffix $s_a$ is said to co-occur with suffix $s_b$ if there is at least one word in the available evidence which occur with both these suffixes. Then for each suffix we can prepare the co-occurrence list, i.e., the list of suffixes that co-occur with it. Let each suffix occur in its own co-occurrence list. If the corpus is large enough, then we can expect that the co-occurrence lists of all suffixes will be exhaustive. For example, in case of the language $L$, the exhaustive co-occurrence list of suffix $a$ is $\{a, d, e, f\}$. This is not to say that there will be necessarily some word in the corpus that occur with all the suffixes $a$, $d$, $e$ and $f$.

Now, let us define a pivot suffix as a suffix that applies to words of a distinct linguistic category. For example, suppose in English if a word takes the suffix -ed, then it is definitely a verb; then ed is a pivot suffix for the class verb. Here, we observe that there can be suffixes that apply to more than one distinct categories of words and hence are not pivot suffixes. For language $L$ of figure 1 $a$, $b$, $c$ and $h$ (or $i$) are pivot suffixes. So if a word occurs with any of these pivot suffixes, then we can tell for sure that the word belongs to the linguistic category represented by that pivot suffix. But in the problem that we have undertaken to tackle, we do not know which of the suffixes are pivot suffixes, or for that matter, how many categories of words and corresponding pivot suffixes are there in the language. Rather finding the list of pivot suffixes is the immediate problem.

In order to find a way to identify the pivot suffixes, first of all consider the suffixes in each linguistic category (refer to figure 1). While the pivot suffixes will figure in only one linguistic category each, the rest will occur in at least two such categories. We claim that the among the suffixes of a given linguistic category, the pivot suffix will have the least number of co-occurring suffixes, and this number is exactly the number of suffixes in that linguistic category. This is easy to justify because all the non-pivot suffixes will occur with at least one suffix from another category, in addition to the suffixes of the category that we are considering. Thus among all the suffixes in the language, the one with the least number of co-occurring suffix will be a pivot suffix. This immediately implies that all the suffixes in the co-occurrence list of this suffix may not be considered as pivot suffixes. To find another pivot suffix we simply have to apply the same criteria (of lowest number of co-occurring suffixes) to the list of suffixes minus the already identified pivot suffix and its co-occurring suffixes. Proceeding like this we can identify all the pivot suffixes. At any stage if we find more than one suffix having the least number of co-occurring suffixes, then we may select any one of them as pivot and proceed as usual. For the language of Figure 1 we shall end up identifying $a$, $b$, $c$ and either $h$ or $i$ as pivot suffixes.

We carry out the classification of words in two broad phases - first, we identify the pivot suffix for each target class, and the co-occurrence list for each pivot suffix. In the second phase we consider each word and their associated suffixes and classify the words as belonging to one (or more) of the classes. If a word occurs with a pivot suffix its classification will be definite. Otherwise its class may be
predicted. Predicted classification means in the given corpus the word has not occurred with adequate number of suffixes so as to tell its class with certainty. In case of predicted classification, a word will be classified into multiple classes. Referring to figure 1 again, we may like to accord both $h$ and $i$ pivot status corresponding to their group for the purpose of classification. Let us call $h$ and $i$ co-pivots. We can accord co-pivot status to those non-pivot suffixes which occur in the co-occurrence list of only one pivot suffix.

The word classes identified above may not correspond to POS classes (tagset). In fact, POS identification (tagging) for words may not be correct unless the context of each occurrence of the words are considered. Our method does not consider the contexts of occurrence of the words. The classes identified in our method can be the linguistic classes that may be entered as attributes of the words in a lexicon. Thus this method may be considered as a tool for developing a lexicon, by providing class information for the entries.

The above steps may be summarised as -

**Phase 1**

1. Form a list of all suffixes in the evidence, $S = \{s_1, s_2, \ldots s_m\}$.
2. For each suffix $s_i$ in $S$, prepare the co-occurrence list, $C_i \subset S$. Note that $s_i \in C_i$.
3. Sort the suffixes by increasing number of elements in the co-occurrence lists.
4. Mark the first suffix in the sorted list as the first pivot suffix $p_1$.
5. Successively from the sorted list of suffixes mark a suffix as pivot suffix if its co-occurrence list does not contain any of the already identified pivot suffixes.
6. For each suffix in the co-occurrence list of each pivot suffix, check if it occurs in the co-occurrence list of any other pivot suffix. If it does not, then mark it as a co-pivot of the pivot suffix.

**Phase 2**

7. For each word, $w_i$ in $W$, if there is a pivot or co-pivot suffix that associated to it, then classify the word in the class represented by that pivot suffix. This classification is definite. Else,

8. If there is some associated suffix of the word that is in the co-occurrence list of a pivot suffix, then classify the word into the class represented by the pivot suffix. This classification is only predictive, i.e., in the absence of any pivot suffix associated with the word, we simply predict that the word can eventually occur with the pivot suffix. For a given word there may be more than one predictive classifications.

**Performance Analysis**  It can be seen that if there are $m$ input words and $n$ suffixes in the language then the complexity of phase 1 is $O(mn + n^2)$ and that of phase 2 is $O(mn)$. Since $m$ is expected to be larger than $n$ (and comparable to $n^2$), hence the overall complexity of the process is $O(mn)$. 


5 Theoretical weaknesses of the model

We need to state here that the above discussion does not cover all possibilities with the set-theoretic model that we have prepared and used. Consider the situation for a hypothetical language $L'$ as shown in Figure 2. Compared to language $L$, in $L'$ $c$ does not exist, there is another suffix $j$ which has $b$ and $g$ as co-occurring suffixes, and there is a category of words that take only suffix $f$. There is no single pivot suffix for the categories corresponding to ellipse 3 and 6, and our above algorithm will fail to detect these categories. In fact we shall get the pivot suffixes $a$, $b$ and $h$ and their co-occurrence lists \{ $a$, $d$, $e$, $f$ \}, \{ $b$, $f$, $e$, $j$ \} and \{ $h$, $g$, $i$ \}.

![Figure 2: Suffixes and linguistic categories of words for language $L'$](image)

Some other such complex situations are depicted in Figure 3.

![Figure 3: Complex co-ocurrences](image)

Though at the cost of simplicity, we can extend our algorithm to cover situations as depicted in
Figure 2 and 3, we feel that most such complexities are only theoretical possibilities and rare in real languages.

### 6 Experimental results

We have carried out the experiments on an Assamese text corpus. Assamese is a Indo-European language of the Indic branch. The grammar, morphology and vocabulary of Assamese can be found in [3], [9], [11], [2], etc. Assamese displays extensive inflection - more than some other Indic languages such as Hindi ([20]). Use of sequences of multiple suffixes in words is also common in Assamese. In our experiment we have limited our analysis to words with single suffixes, though some such words may also have occurred with more suffixes in sequence. For example, suppose the corpus contains the words colour-ed and colour-ed-ness, then we consider colour-ed but not colour-ed-ness, and try to determine the class for colour.

In our experiment the input corpus was of about 49000 words (of which about 11500 were distinct). These words were segmented into base words and suffixes using an unsupervised morphology acquisition method described in [18]. Then for the purpose of this classification we selected only those words that were not further decomposed (base words) and occurred with regular suffixes (i.e., suffixes that have occurred with several words). We also left out the cases of words attaching with other words to form compound words (this selection was also done using an unsupervised method). Thus we had about 800 words on which we tried this classification method. The summary of the results is -

- **Number of suffixes - 38**
- **Pivot suffixes and their co-occurrence lists (co-pivots are underlined)** -

<table>
<thead>
<tr>
<th>Pivot Suffix</th>
<th>Co-occurrence list</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>bhAwe</td>
<td>bhAwe, kE, tA</td>
<td>bhISN, gbIR,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mOv, mrNO/trr</td>
</tr>
<tr>
<td>bhae</td>
<td>bhae, b, e</td>
<td></td>
</tr>
<tr>
<td>chil</td>
<td>t, o, ye, n, chil, che, b, l, He</td>
<td>AgbrhAI, AorAI</td>
</tr>
<tr>
<td>chil</td>
<td></td>
<td>HE, shunAI, uThi</td>
</tr>
<tr>
<td>ai</td>
<td>r, t, k, e, lE, To, I, ai, khn, skle</td>
<td>b_/Hm, chAT_/r</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sh/bd</td>
</tr>
<tr>
<td>sh</td>
<td>r, t, o, To, n, ere, Tor, it, ly, znk</td>
<td>AHt, dkh,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hAbh</td>
</tr>
<tr>
<td>re</td>
<td>t, i, o, n, khn, b, zn, ke, ei, bor, re</td>
<td>dA, gch, ghRhNA,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ju/kti, p_/_rJek</td>
</tr>
<tr>
<td>smUH</td>
<td>r, t, k, e, i, o, ye, To, A, I, Tor, smUH</td>
<td>gocr, guN,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sh/kti, sh/str</td>
</tr>
<tr>
<td>sOuH</td>
<td>sOuH, b, s, s, s, s, ke, s, ke, s, ke</td>
<td></td>
</tr>
<tr>
<td>te</td>
<td>r, e, i, lE, To, khn, zn, l, Tor, ei, te, it, grAKI</td>
<td>bRh/tti, bh/gmN,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mU, mUTh, pAt</td>
</tr>
<tr>
<td>TA</td>
<td>r, t, k, e, i, o, ye, n, khn, r, zn, l, ke, TA, grAKI</td>
<td>Ashl, cAri,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dH, duyo</td>
</tr>
</tbody>
</table>

- **Classifications** -
115 definite classifications,
1988 predictive classifications where all suffixes of each word are present in the co-occurrence group of the pivot suffix,
1354 predictive classifications where some suffixes of each word are not present in the co-occurrence group of the pivot suffix.

In Assamese a majority of the words are used as nouns. The proportion of verbs and adjectives is low. The high point in our results is that the pivot suffix bhAwe represents the class of adjectives and adverbs very well, and similarly the pivot suffix chil represents the class of verbs. The other pivot suffixes represent nouns but with some differences, e.g., TA represents the class of numbers and other quantifiers that are used as nouns. However, we observe two kinds of impreciseness in our experimental results - firstly, some definite classifications are not correct. For example, g/lp and pAb (পণ, পর) are put in the same category due to the suffix ai (অই), though the former is a noun and the latter is a verb. Secondly, not all the categories of nouns are significant, few can be avoided, e.g., the category represented by ai (অই) can probably be avoided. The primary reason for such imprecision is the small size of the evidence that we have used compared to the number of suffixes. Also, in few of the words the segmentations as obtained by an unsupervised method were not exact, e.g. pAb should have been considered as pA + b.

7 Shortcomings in the Pivot Suffix Model of Classification

It can be seen that the above unsupervised method of identifying word categories depends heavily on co-occurrences of suffixes, and hence, will club together linguistic categories of words if there are words that fall in multiple linguistic categories. This is because the co-occurrence evidence of such ambiguous words will lead to one (or more) pivot suffix to be eclipsed by another, and thereby spurious merging of multiple word categories. We feel that either some statistical figures may be taken into account to reduce such influences, or the unsupervised identification of classes should be tried using simple corpus that does not contain such ambiguities. Alternatively, we can resort to supervision in the form of externally identifying the vulnerable pivot suffixes.

Another shortcoming of the whole idea of word classification based on suffix evidence is that there can be some categories of words that do not take any suffixes. For example, in Assamese, the words Aru, bA, ki/ntu, (আক, ব, কিস্ত, meaning and, or, but), etc. generally do not take any suffix. This method may at most identify all such words as belonging to a single category. For a highly inflectional language like Assamese, this is not a very serious problem, however (all the above three words are conjunctions). However, one should be careful in such decisions because a word may appear without suffix only in the corpus being considered. A simple criteria can be - if the word has occurred a large number of times in the corpus, but always occurred without suffix, then probably it never takes any suffix.
8 Conclusion

In this work we have represented the problem of affix-based word classification in terms of sets and developed a solution for unsupervised identification of linguistic categories of words in a language using information of association of suffixes with different words in a corpus. Then we have described how, the words can be subsequently classified into the identified categories using the same input morphological evidence.

References


