CS671 – Natural Language Processing

HW1

~ Ankit Modi (10104)

Undivided++ was used for the mentioned morphological analysis. *Hi_Blogs.txt* file was used to train the morphological analyzer.
First of all, words and their frequencies were extracted from the above mentioned file using sed and awk. The output file generated had frequency of each word followed by the word itself (separated by a space) in each of its lines.
Initially, around 0.3 million distinguished words were generated. But that list had many repetitions such as 'अधिकार', 'अधिकार,', 'अधिकार..', 'अधिकार|'.
So, all the instances of comma(,), full stop(.), hyphen(-) etc were changed to space (space was being used as the delimiter) which resulted in removing this redundancy.
All the words having 3 or less than 3 Unicode characters were removed. Also, since we are working with Hindi dataset so all the occurrences of English words (links etc) and numerical values were removed from the list.
All these pruning methods resulted in reducing the list to around 0.21 million distinguished words which took less run time on local machine.

Following parameters were used:
#define SMALL_ROOT_LENGTH 3

#define LOW_FREQUENCY_DROPOUTS 1

#define LOW_FREQUENCY_DROPOUTS_LEARNING 5

#define SUFFIX_CUTOFF_THRESHOLD 50

#define PREFIX_CUTOFF_THRESHOLD 70

#define COMPOSITE_SUFFIX_THRESHOLD 0.65

#define WRFR_SUFFIX_THRESHOLD 10

#define WRFR_PREFIX_THRESHOLD 1.5

#define SLS_NORMALIZATION_CONSTANT 7

#define ALLOMORPH_REPLACEMENT_THRESHOLD 4

#define ALLOMORPH_DELETION_THRESHOLD 4

#define ALLOMORPH_ADDITION_THRESHOLD 4

#define PROMOTE_LONG_SEGMENTATION 1

#define PROMOTE_LONG_SEGMENTATION_LENGTH 15

#define INDUCE_OUTOFVOCABULARY_ROOTS 1

#define INDUCE_OUTOFVOCABULARY_ROOTS_THRESHOLD 5

Table consists of undivided++ parameters used at run time and respective values of precision, recall and F_score.

Undivided ++ Parameters	Precision	Recall	F_score
1 0 0 1 0	0.48	0.33	0.40

Errors occurring in the morphemes generated by undivided++ were mainly due to two facts:

1. Data set didn't include all the morphemes of the words present in the test set. For example, अधिकारियों was broken into अधिक and अरियों as अधिक was present in the training set and not अधिकार, which should be the correct morpheme.
2. Some of the words were of English origin so their morphemes were not present in the data set which resulted in their identification as one single morpheme.