Volume 5, Issue 1, January 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Number Tagger for Hindi Morphological Analyzer

Kanak Mohnot Arora[1], Vipin Singh[2], Mahaveer Sain[3]

Abstract - Number Tagger is an important tool that is used to develop Part of Speech Tagger and morphological analyzer. The problem of tagging in natural language processing is to find a way to tag every word in a text as a particular number. In this paper, we present a Number Tagger for Hindi. Our System is evaluated over a corpus of 80,000 words with 2 different standard number (i.e. Singular and Plural) tags for Hindi. Accuracy is the prime factor in evaluating any tagger so the accuracy of proposed tagger is also discussed in this paper.

Keywords-Number, Tagging, Rules.

I. INTRODUCTION

A word can be defined as a sequence of characters separated by spaces, punctuation marks, etc. in case of the written text. A word can be of two types: simple and compound. A simple word or word consists of a root or stem together with suffixes and prefixes. A compound word can be broken into two or more independent words. Each of the constituent words in a compound word is either a compound word or a simple word and may be used independently as a word. On the other hand, the root and the affixes, which are constituents of a simple word, are not independent words and cannot occur as separate words in the text. Constituents of a simple word are called morphemes or meaning units. The overall meaning of a simple word comes from the morphemes and their relationships [1]. Morphological Analysis is the process of finding the constituent morphemes in a word like बिल्ली -N +PL for word बिल्लियाँ [2]. Morphological generator is the process of generating the word form taking stem word and its features (affixes) as input. Morphological Analysis is essential for Hindi it has a rich system of inflectional morphology as like other Indo-Aryan family languages. Main concern here is on the grammatical information of words and this grammatical information like gender, number, person etc. is marked through the inflectional suffixes. [5][6]

Number tagger is an important application for developing part of speech tagger. It is an important part of morphological analyzer. It is the process of assigning a number like singular or plural to each word in a sentence. [3][4]

II. SYSTEM DESCRIPTION

This block is used to identify the number of the input tokens. The number of the token is found by applying various types of rules.

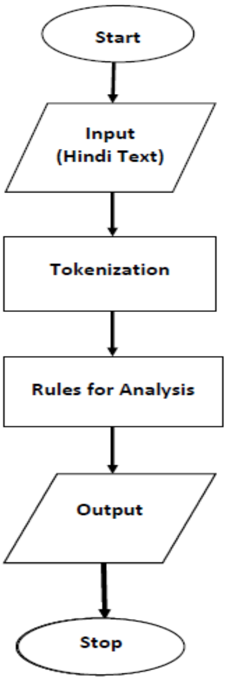

Figure 1:- Number Flow Chart

A. Algorithm

Steps:

1. Enter the Hindi input text.
2. Tokenize the input text i.e. break the input sentence (according to phrases and single tokens) into independent and meaningful words.
3. Find the number of the token by applying number rules.
4. Display the result on the screen.

B. Following Rules are applied to identify different Tags

Rule 1:-The token ending with the suffix “एँ“,”यों”,”औं” etc. are termed as plural token.

e.g. माताएँ, राजों केले

Rule 2:-The token having a quantity before it are termed as plural token.

e.g. चार आम, एक किलो सेब

III. EVALUATION AND RESULT

Table 1: Test Cases

Sentence	**Number**
आइए जानतेहैं दिल्ली में हुए इस उलटफेर की कुछ वजहें	आइए-एकवचन / जानतेहैं-एकवचन / दिल्ली-एकवचन / में-एकवचन / हुए-एकवचन / इस-एकवचन / उलटफेर-एकवचन / की-एकवचन / कुछ-एकवचन / वजहें-बहुवचन
केंद्र सरकार के खिलाफ आम लोगों का गुस्सा शीला सरकार को झेलना पडा	केंद्र सरकार-एकवचन / के-एकवचन / खिलाफ-एकवचन / आम लोगों-बहुवचन / का-एकवचन / गुस्सा-एकवचन / शीला-एकवचन / सरकार-एकवचन / को-एकवचन / झेलना-एकवचन / पडा-एकवचन
दोपहर आते आते ढोल नगाड़े यहां पहुंच गए और समर्थकनाच गाकर जश्न मनाने लगे	दोपहर-एकवचन / आते आते-एकवचन / ढोल नगाड़े-बहुवचन / यहां-एकवचन/ पहुंच-एकवचन / गए-एकवचन / और-एकवचन / समर्थकनाच-एकवचन / गाकर-एकवचन / जश्न-एकवचन / मनाने-बहुवचन / लगे-एकवचन

The evaluation metrics for the data set is precision, recall and F-Measure. These are defined as following:-

A. Precision

Precision is the ratio of the number of items of a certain named entity type correctly identified to all items that were assigned that particular type by the system.

$$P = \frac{\text{Number of correct tags assigned}}{\text{Total number of tags assigned}}$$

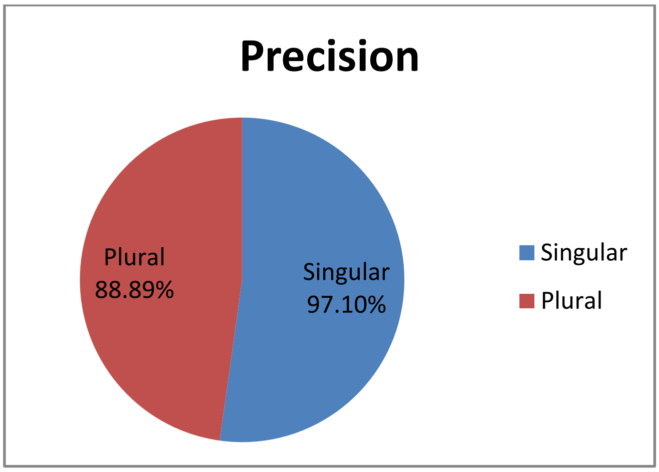

Figure 2: Precision Pie Chart

B. Recall

Recall measures the number of items of a certain named entity type correctly identified, divided by the total number of items of this type. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

$$R = \frac{\text{Number of correct tags assigned}}{\text{Total number of tags in the annotated test corpus}}$$

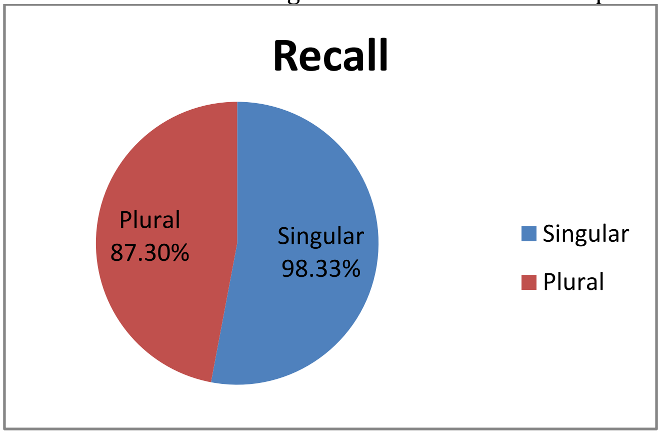

Figure 3: Recall Pie Chart

C. F-measure

The F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision P and the recall R of the test to compute the score. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0. It combines Recall (R) and Precision (P) using the formula .The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall:

$$F = \frac{2\mathrm{RP}}{\mathrm{R} + \mathrm{P}}$$

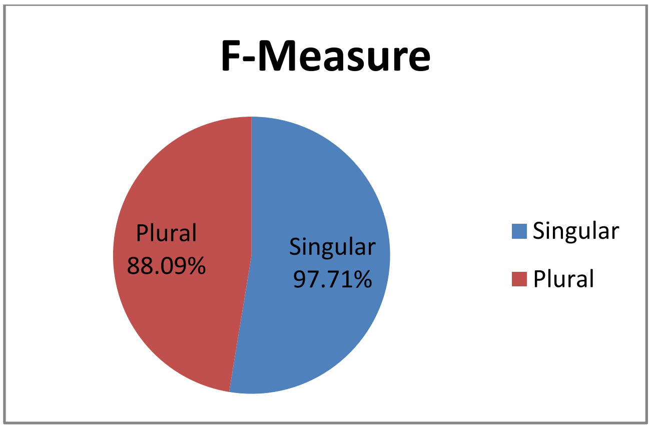

Figure 4: F-Measure Pie Chart

IV. CONCLUSION

At last we conclude that Number tagging is the most important activity Part Of Speech tagger. The accuracy of any POS tool is dependent on the accuracy of singular formation. Different approaches have been used by authors for the development of part of speech tagger for Indian Languages.

We have presented a number tagger for Hindi which is used by our Part of Speech tagger. We have shown that such a system has good performance with an average accuracy of 93.9% for number tagging. We believe that further error analysis and more language specific features would improve the system performance.

REFERENCES

[1] Bharati, Akshar, Vineet Chaitanya and Rajeev Sangal. (1995). Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi.

[2] Bharati, Akshar, Amba P. Kulkarni, Vineet Chaitanya. (1998a). Challenges in Developing Word Analyzers for Indian Languages, Presented at Workshop on Morphology, CIEFL, Hyderabad, July 1998.

[3] LTRC, IIIT Hyderabad http://ltrc.iiit.ac.in.

[4] Aduriz 1., Agirre E., 'A word-grammar based morphological analyzer for agglutinative languages'**,** University of lhe Basqtlo Cotlnlry, Basque Country.

[5] Uma Parameshwari Rao G, Parameshwari K: CALTS, University of Hyderabad, 'On the description of morphological data for morphological analyzers and generators: A case of Telugu, Tamil and Kannada'.

[6] Beesley, K. and L. Karttunen. 'Finite State Morphology'. Stanford, CA: CSLI Publications, 2003.