Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge

Ms.Ashwini VairalkarMs.Amita Mahajn Ms.Darshana Sonawane

[1]*Student, Department of Information Technology, G.H.Raisoni Institute of Engineering andmanagement, M.S,India.*
[2]*Student, Department of Information Technology, G.H.Raisoni Institute of Engineering andmanagement, M.S,India.*
[3]*Student, Department of Information Technology, G.H.Raisoni Institute of Engineering andmanagement, M.S,India.*

ABSTRACT

Part of Speech tagging for Indian Lan- guages in general and Hindi in particu- lar is not a very widely explored territory. There have been many attempts at devel- oping a good POS tagger for Hindi, but the morphological complexity of the language makes it a hard nut to crack. Some of the best taggers available for Indian Lan- guages employ hybrids of machine learn- ing or stochastic methods and linguistic knowledge. Though, the results achieved using such methods are good, there prac- ticability for other inflective Indian Lan- guages is reduced due to their heavy de- pendence on linguistic knowledge. Even though taggers can achieve very good re- sults if provided good morphological in- formation, the cost of creating these re- sources renders such methods impractical.

1. INTRODUCTION

Marine pollution occurs when harmful effects result from the entry into the ocean of chemicals, ((particles, industrial, agricultural and residential waste, noise, or the spread of invasive organisms. Eighty percent of marine pollution comes from land. Air pollution is also a contributing factor by carrying off pesticides or dirt into the ocean. Land and air pollution have proven to be harmful to marine life and its habitats.

The pollution often comes from nonpoint sources such as agricultural runoff, wind-blown debris, and dust. Pollution in large bodies of water can be aggravated by physical phenomena like wind driven Langmuir circulation and their biological effects. Nutrient pollution, a form of water pollution, refers to contamination by excessive inputs of nutrients. It is a primary cause of eutrophication of surface waters, in which excess nutrients, usually nitrates or phosphates, stimulate algae growth. Many potentially toxic chemicals adhere to tiny particles which are then taken up by plankton and benthic animals, most of which are either deposit feeders or filter feeders. In this way, the toxins are concentrated upward within ocean food chains. Many particles combine chemically in a manner highly depletive of oxygen, causing estuaries to become ((anoxic.

When pesticides are incorporated into the marine ecosystem, they quickly become absorbed into marine food webs. Once in the food webs, these pesticides can cause mutations, as well as diseases, which can be harmful to humans as well as the entire food web. Toxic metals can also be introduced into marine food webs. These can cause a change to tissue matter, biochemistry, behavior, reproduction, and suppress growth in marine life. Also, many animal feeds have a high fish meal or fish hydrolysate content. In this way, marine toxins can be transferred to land animals, and appear later in meat and dairy products.

SMOOTHING

The implementation of the proposed system required cautious handling of small numbers and zero probabilities at various points. First, propagating and multiplying partial probabilities in the induction step of Viterbi led to number underflows. This issue can be handled during training: In order to eliminate this problem, I used the natural logs of the transition and emission probabilities, and changed the related multiplications into summations. Note that this problem does not apply to the other two algorithms that were implemented, because both of them disregard partial probabilities

Posting

Part-of-Speech (POS) tagging is the process of automatic annotation of lexical categories. Part-of–Speech tagging assigns an appropriate part of speech tag for each word in a sentence of a natural language. The development of an automatic POS tagger requires either a comprehensive set of linguistically motivated rules or a large annotated corpus. But such rules and corpora have been developed for a few languages like English and some other languages. POS taggers for Indian languages are not readily available due to lack of such rules and large annotated corpora. A part-of-speech is a grammatical category commonly including nouns, pronouns, verbs, adjectives, adverbs, prepositions, conjunctions, interjections. Parts of speech can be divided into two broad categories: closed classes and open classes.

2. CONCLUSIONS

The over all performance of this approach is bet- ter than a simple stochastic method. But, it can- not hold a candle to methods using detailed mor- phological analysis and linguistic resources. The results presented may not be very impressive if compared to methods similar to one presented in (Singh et al., 2006), but, they prove that a simple stochastic method can be easily modified and used for improving performance by harnessing mor- phology in the simplest manner possible. In this paper, our aim was to demonstrate a method which can give good performance without relying on ex- tensive linguistic knowledge.

3. REFERENCE

1. Doug Cutting, Julian Kupiec, Jan Pedersen, and Pene- lope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.

2. S. Connell. 1996. A comparison of hidden markov model features for the recognition of cursive hand- writing, May. Master's Thesis, Dept. of Computer Science, Michigan State University.

3. Kevin Duh and Katarin Kirchoff. 2004. Pos tagging of dialectal arabic: A minimally supervised approach
.

4. Ezra Black, Fred Jelinek, John Lafferty, Robert Mer- cer, and Salim Roukos. 1992. Decision tree models applied to the labeling of text with parts-of-speech. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 117–121, Morristown, NJ, USA. Association for Computational Linguis- tics.

5. Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1995. *Natural Language Processing – A Paninian Perspective*. Prentice-Hall Indian.