Mini Project - POS Taggers for Indian Languages

INTRODUCTION:

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language. One of the fundamental tasks in NLP is Part-of-Speech (POS) tagging, which involves assigning grammatical categories (such as nouns, verbs, adjectives, etc.) to words in a sentence. POS tagging plays a crucial role in various NLP applications, including machine translation, named entity recognition, sentiment analysis, and information retrieval.

While POS tagging is well-established for languages like English, it presents unique challenges for Indian languages such as Hindi, Bengali, Tamil, etc. Indian languages are characterized by their morphological complexity, rich linguistic diversity, and lack of standardized resources compared to English. As a result, developing accurate and reliable POS taggers for Indian languages is a non-trivial task that requires careful consideration of linguistic phenomena specific to each language.

The objective of this mini project is to develop a POS tagger for Hindi, one of the most widely spoken languages in India. Hindi is an Indo-Aryan language spoken by millions of people in India and other parts of the world. Despite its prevalence, the development of NLP tools and resources for Hindi lags behind languages like English, posing challenges for NLP research and applications in Hindi.

By developing a POS tagger for Hindi, this mini project aims to address this gap and contribute to the advancement of NLP research and technology in Indian languages. A robust POS tagger for Hindi can facilitate various NLP tasks, including text analysis, information extraction, document classification, and more, thereby enabling the development of innovative applications tailored to Hindi-speaking users.

Furthermore, this mini project serves as a stepping stone towards building a deeper understanding of linguistic structures and patterns in Hindi text. Through the development and evaluation of the POS tagger, we gain insights into the grammatical complexities of Hindi, including verb conjugation, noun declension, adjective agreement, and more. This knowledge

can inform future research in Hindi linguistics and computational linguistics, driving innovation and progress in the field.

METHODOLOGY

Dataset:

The dataset used for this project is a corpus of Hindi text annotated with POS tags. The dataset, named "hindi.pos," contains sentences in Hindi along with their corresponding POS tags. This annotated dataset serves as the training data for building and evaluating the POS tagger model.

POS Tagger:

Part-of-Speech (POS) tagging, also known as grammatical tagging or word-category disambiguation, is a fundamental task in Natural Language Processing (NLP). It involves assigning grammatical categories, or "tags," to words in a sentence based on their syntactic and morphological properties. These tags typically represent the word's part of speech, such as noun, verb, adjective, adverb, pronoun, preposition, conjunction, etc.

Importance of POS Tagging:

POS tagging plays a crucial role in various NLP applications, including:

Text Analysis: POS tagging helps in extracting meaningful information from text by identifying the grammatical structure of sentences. This information can be used for tasks such as parsing, semantic analysis, and information extraction.

Named Entity Recognition (NER): POS tags can be used as features in NER systems to identify and classify named entities such as persons, organizations, locations, dates, etc., within text.

Machine Translation: POS tags provide linguistic information that can aid in translating sentences from one language to another by preserving the grammatical structure and semantic relationships.

Information Retrieval: POS tags can be used to improve the accuracy of search engines by enabling more precise query matching based on syntactic patterns.

Sentiment Analysis: POS tagging can be used as a feature in sentiment analysis models to analyze the sentiment expressed in text based on the parts of speech used.

POS Tagging Process:

The process of POS tagging involves the following steps:

1. Tokenization: The input text is segmented into individual tokens, which typically correspond to words or punctuation marks. Tokenization is essential for identifying the boundaries between words in the text.
2. Lexical Analysis: Each token is mapped to its corresponding lemma or base form. This step helps in reducing the complexity of the tagging process by standardizing the representation of words.
3. POS Tagging: Each token is assigned a POS tag based on its context within the sentence and its morphological properties. POS tagging can be done using rule-based approaches, statistical models, or hybrid methods that combine both approaches.
4. Evaluation: The accuracy of the POS tagger is evaluated using benchmark datasets with manually annotated POS tags. Common evaluation metrics include accuracy, precision, recall, and F1 score.

POS Tagging Techniques:

1. Rule-Based Approaches: Rule-based POS taggers rely on handcrafted linguistic rules to assign POS tags to words based on their context and syntactic patterns. These rules are often derived from linguistic theories and grammatical conventions.
2. Statistical Approaches: Statistical POS taggers learn POS tag patterns from annotated corpora using machine learning algorithms such as Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), Conditional Random Fields (CRFs), and Neural Networks. These models

estimate the probability of observing a particular POS tag given the current word and its context.

3. Hybrid Approaches: Hybrid POS taggers combine rule-based and statistical techniques to leverage the strengths of both approaches. For example, rule-based patterns can be used to improve the accuracy of statistical models, while statistical models can handle unseen or ambiguous cases better than rule-based systems.

RESULTS:

In [1]:

```
import nltk
from nltk.corpus import indian
from nltk.tag import tnt
import string


nltk.download('punkt')
#nltk.download()

tagged_set = 'hindi.pos'
word_set = indian.sents(tagged_set)
count = 0
for sen in word_set:
    count = count + 1
    sen = "".join([" "+i if not i.startswith("'") and i not in string.punctuation else i for i in sen]).strip()
    print (count, sen)
print ('Total sentences in the tagged file are',count)

train_perc = .9

train_rows = int(train_perc*count)
test_rows = train_rows + 1

print ('Sentences to be trained',train_rows, 'Sentences to be tested against',test_rows)

data = indian.tagged_sents(tagged_set)
train_data = data[:train_rows]
test_data = data[test_rows:]


pos_tagger = tnt.TnT()
pos_tagger.train(train_data)
pos_tagger.evaluate(test_data)



sentence_to_be_tagged = "३९ गेंदों में दो चौकों और एक छक्के की मदद से ३४ रन बनाने वाले परोरे अंत तक आउट नहीं हुए ।"

tokenized = nltk.word_tokenize(sentence_to_be_tagged)
```

```
print(pos_tagger.tag(tokenized))
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\PREDATOR/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
C:\Users\PREDATOR\AppData\Local\Temp\ipykernel_15796\1723484861.py:38: DeprecationWarning:
  Function evaluate() has been deprecated.  Use accuracy(gold)
  instead.
  pos_tagger.evaluate(test_data)
```

1 पूर्ण प्रतिबंध हटाओ: इराक
2 संयुक्त राष्ट्र ।
3 इराक के विदेश मंत्री ने अमरीका के उस प्रस्ताव का मजाक उड़ाया है, जिसमें अमरीका ने संयुक्त राष्ट्र के प्रतिबंधों को इराकी नागरिकों के लिए कम हानिकारक बनाने के लिए कहा है ।
4 विदेश मंत्री का कहना है कि चूंकि बगदाद संयुक्त राष्ट्र की मांगों का पालन करते हुए अपने भारी विनाशकारी हथियारों को नष्ट कर रहा है ।
5 लिहाजा प्रतिबंधों को पूर्ण रूप से उठा दिया जाना चाहिए ।
6 विदेश मंत्री मोहम्मद सईद का कहना है कि वे इसे'सुव्यवस्थित प्रतिबंध' कह कर आम राय और सुरक्षा परिषद को छल रहे हैं ।
7 बेनजीर की सुनवाई स्थगित
8 कराची ।
9 पाकिस्तान की पूर्व प्रधानमंत्री बेनजीर भुट्टो पर लगे भ्रष्टाचार के आरोपों के खिलाफ भुट्टो द्वारा दायर की गई याचिका की सुनवाई मंगलवार को वकीलों की हड़ताल के कारण स्थगित कर दी गई ।
10 सिंध हाईकोर्ट बार एसोसिएशन के अध्यक्ष रशीद रिजवी के मुताबिक यह हड़ताल उच्च न्यायालय और निचली अदालतों के स्तर पर सफल रही ।
11 देश में पुन: प्रजातंत्र की स्थापना की मांग को लेकर यह हड़ताल की गई थी ।
12 सुप्रीम कोर्ट में भुट्टो के उक्त मामले की सुनवाई सोमवार से शुरू हुई, जो फिलहाल बुधवार तक स्थगित है ।
13 मुशर्रफ सऊदी अरब को मनाएंगे
14 इस्लामाबाद ।
15 पाकिस्तानी सैन्य प्रशासक जनरल परवेज मुशर्रफ द्वारा सऊदी अरब को कश्मीर विवाद मुद्दे पर भारत और पाकिस्तान की मध्यस्थता के लिए मनाने की संभावना व्यक्त की गई है ।
16 आधिकारिक सूत्रों के हवाले से कहा गया है कि मुशर्रफ की बुधवार से शुरू हो रही हज यात्रा इस मायने में काफी महत्वपूर्ण हो सकती है ।
17 मुशर्रफ अपने इस हज यात्रा के दौरान कश्मीर के अलावा फिलीस्तिन, अमरीका द्वारा ओसमा बिन लादेन के प्रत्यर्पण की मांग आदि पर चर्चा करेंगे ।
18 इस दौरान मुशर्रफ सऊदी राजा फाहद, उप प्रधानमंत्री सहित कई बड़ी हस्तियों से मिलेंगे ।
19 पाकिस्तान के बर्खास्त पूर्व प्रधानमंत्री नवाज शरीफ के मामले में भी यहां चर्चा होगी ।
20 पत्रकारों के लिए कल्याण कोष
21 नई दिल्ली ।
22 वित्तमंत्री यशवंत सिन्हा ने अपने बजट में पत्रकारों के लिए कल्याण कोष स्थापित करने की घोषणा की है ।
23 कोष में सूचना एवं प्रसारण मंत्रालय १ करोड़ की सहायता देगा ।
24 बजट प्रस्तुत करते हुए सिन्हा ने कहा कि पत्रकार कई अवसरों पर कवरेज के लिए जोखिम उठाते हैं यह कोष उनकी सेवा एवं संघर्ष के लिए होगा ।
25 उन्होंने कहा कि पांच वर्ष में एक बार मान्यता प्राप्त पत्रकार को करमुक्त आयात द्वारा व्यक्तिगत उपयोग के लिए सामग्री मंगाने की सुविधा अब दो वर्ष कर दी गई है ।
26 फिलहाल मान्यता प्राप्त पत्रकार एवं कैमरामैन कैमरा, फैक्स, कम्प्यूटर आदि जिनकी अधिकतम लागत १ लाख रुपए तक हो बिना कस्टम ड्यूटी चुकाए पांच साल में एक बार मंगा सकते थे ।
27 बजट विकास को गति देगा: वाजपेयी