Syntax Analysis

Date
Page

Parts of speech tagging:

i] Parts of speech tagging is the process of assigning a POS noun, verb, pronoun, preposition, adverb and adjective to each word in a sentence.

ii] The input to a tagging algorithm is the sequence of words of a natural language sentence and specified tag sets. The output is a single best part-of-speech tag for each word. Many words may belong to more than one lexical category.

iii] For example, English word 'book' can be noun as in "I am reading a good book" or a verb as in "The police booked the snatcher".

iv] The same is true for other languages. for example, the Hindi word 'sona' may mean 'gold' (noun) or 'sleep' (verb). However, only one of the possible meaning is used at a time.

v] In tagging, we try to determine the correct lexical category of a word in its context. No tagger is efficient enough to identify the correct lexical category of each word in a sentence in every case.

vi] The tag assigned by a tagger is the most likely for a particular use of word in a sentence. The collection of tags used by a particular tagger is called a tag set.

vii] Most part-of-speech tag sets make use of same basic categories, i.e. noun, verb, adjective, prepositions. However, tag sets differ in how they define categori and how finely they divide words into categories.

viii] In addition, most tag sets capture morpho-syntactic information such as singular/plural, number, gender, tense, etc.

ix] consider the following sentences:-
Zuha eats an apple daily
Aman ate an apple yesterday
They have eaten all the apples in basket
I like to eat guavas
The word 'eat' has a distinct grammatical form in each of these four sentence. Eat is the base form, ate is it's past tense, and the form eats requires third pers singular subject. Similarly, eaten is the past participle form and cannot occur in another grammatical conte It is required after have or has.

x] It is required Thus, the following sentences are ungramm
I like to eats guava
They eaten all the apples.
The number of tags used by different taggers varies substantially.

xi] Penn Treebank tag set contains 45 tags while c7 uses 164. For a language like English, which is not morphologically rich, c7 tagset is too big.

xii] The tagging process would yield too many mistagged words and the result would have to be manually corrected.

xiii] Despite this, bigger tag sets have been used, eg - Tosca-ICE with 270 tags. The larger the tag set, greater the information captured about linguistic context. However, the task of tagging becomes complicated and requires manual correction.

xiv] A bigger tag set can be used for morphologically rich languages without introducing too many tagging errors. A tag set that uses just one tag to denote all the verbs will assign identical tags to all the forms of a verb.

xv] Although this coarse-grained distinction may be appropriate for some tasks, a fine-grained tag set captures more information. This is useful for tasks like syntactic pattern detection.

xvi] Penn Treebank tag set captures finer distinctions by assigning distinct tags to distinct grammatical forms of a verb, as summarized below:-

VB	Verb, base form Subsumes imperative, infinitives
VBD	Verb, past tense Includes conditional form of verb "to be"
VBG	Verb, gerund, or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

xvii] Tags assigned to the four different forms of the word 'eat' according to this tag set is as shown below:-

eat	VB
ate	VBD
eaten	VBN
eats	VBP

xviii] Here is an example of a tagged sentence

Speech/NN sounds/NN were/VBD sampled/VBN by/IN a/DT microphone/NN

The tag set used is Penn Treebank.

Another tagging possible for this sentence is as follows:-

Speech/NN Sounds/VBZ were/VBD sampled/VBN by/IN a/DT microphone/NN.

xix] It is easy to see that the second tagged sequence is not corrected. It leads to semantic incoherence we resolve the ambiguity using context of the word. The context is also utilized by automatic taggers.

xx] Pos tagging is an early stage of text processing in many NLP applications including speech synthesis, machine translation, information retrival and information extraction.

xxi] In information retrieval, pos tagging can be used for indexing and for disambiguating word senses.

Tagging is not as complex as parsing.

xxii] In tagging, a complete parse tree is not built; parts of speech is assigned to words using contextual information.

xxiii] Pos tagging methods fall under three categories -
- Rule-based (linguistic)
- Stochastic (data-driven)
- Hybrid

xxiv] Rule-based taggers use hand-coded rules to assign tags to words. These rules use a lexicon to obtain a list of candidate tags and then use rules to discard incorrect tags.

xxv] Stochastic taggers have data driven approaches in which frequency-based information is automatically derived from corpus and used to tag words.

Stochastic taggers disambiguate words based on the probability that a word occurs with particular tag.

xxvi] Hybrid taggers combine features of both these approaches like rule-based systems, they use rules to specify tags like Stochastic system, they use machine-learning to induce rules from a tagged training corpus automatically Brill tagger is an example of hybrid approach.