Tagging: descriptors to the given tokens is

Life, Friendship

Tagging: The descriptors are called the tags and theautomatic assignment of the descriptors to the given tokens is called tagging. POS TaggingThe process of assigningone of the parts of speech to the given word is called Parts Of Speech tagging, commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categoriesPOS TaggerAPart-Of-Speech Tagger (POS Tagger) is a software that reads text and then assigns partsof speech to each word (and other token), such as noun, verb, adjective, etc., It uses different kinds of information such as dictionary, lexicons, rules, etc. because dictionaries have category or categories of a particular word, that is a word may belong to more than one category. For example, run is bothnoun and verb so to solve this ambiguity taggers use probabilistic information. Thereare mainly two type of taggers: Rule-based- Uses hand-written rules to distinguish the tag ambiguity. Stochastictaggers are either HMM based – chooses the tag sequence which maximizes theproduct of word likelihood and tag sequence probability, or cue-based, usingdecision trees or maximum entropy models to combine probabilistic features.

HMMHiddenMarkov Model (HMM) is a statistical Markov model in which the system beingmodeled is assumed to be a Markov process with unobserved (i. e. hidden) states.

In Markov models, the state is directly visible, and thus the state transition probabilitiesare the only parameters, while in the hidden Markov model, the state is not directlyvisible, but the output dependent on the state is visible. Each state has aprobability distribution over the possible output tokens. Therefore, the sequenceof tokens generated by an HMM gives some

information about the sequence ofstates. Accuracyachieved TheEuropean group developed CLAWS, a tagging program that did exactly this, andachieved accuracy in the 93–95% range. Manymachine learning methods have also been applied to the problem of POS tagging. Methods such as SVM, maximum entropy classifier, perceptron, and nearest-neighborhave all been tried, and most can achieve accuracy above 95%. Amore recent development is using the structure regularization method forpart-of-speech tagging, achieving 97. 36% on the standard benchmark dataset.

TagsetAset of tags from which the tagger choses a relevant tag for the word. Data setAmerged Bhojpuri dataset containing of sentences of Bhojpuri and thecorresponding labels to the words. Natural Language Processing(NLP) withPython NLTKis a leading platform for building Python programs to work with human languagedata.

It provides easy-to-use interfaces to over 50 corpora and lexicalresources such as WordNet, along with a suite of text processing libraries forclassification, tokenization, stemming, tagging, parsing, and semanticreasoning, wrappers for industrial-strength NLP libraries, and an active discussionforum. It has many libraries to work on natural language. Using we can tokenizeand tag some text, identify some named entities and display a sparse tree. ACKNOWLEDGEMENT I express my profound and sincere gratitude to my mentor Dr. Anil KumarSingh for providing me with all the facilities and support during my winterinternship period. I would like to thank my guide Mr. Rajesh Mundotiya for their valuableguidance, constructive criticism, encouragement and also for making the

requisiteguidelines enabling me to complete my work with utmost dedication andefficiency. At last, I would like to acknowledge my family and friends for themotivation, inspiration and support in boosting my moral without which myefforts would have been in vain.