**Abstract**

Most of the work associated to natural language processing NLP has been done in many different languages. For instance, English, German, Chinese, Arabic and Spanish. Many natural language processing (NLP) methods can be used for extracting information from data. These part of speech taggers are very important modules in many applications such as word processors, grammar checkers, speech recognizers etc.

**Introduction**

Many natural language processing (NLP) methods can be used for extracting information from data [**Developing a Robust Part-of-Speech Tagger for Biomedical Text**]. The main objective of this research is to develop a system for Pashto language which can automatically detect the part of speech tag for a specific word. In natural language processing Part of Speech (POS) tagging has a very important role. It is the process of assigning each word a specific part of speech tag POS. There are many parts of speech tags exists but the most commonly used POS tags are Verb, Noun, Adjective and Adverb [**Rule-Based Part of Speech Tagging for Pashto Language**]. These tags can be further divided into sub categories such as verbs can be further classified into past tense future tense as well as nouns can be further classified into singular or plural nouns. There are many important applications available for part of speech POS taggers, for instance machine translation, speech recognition, information retrieval, speech synthesis and many more [**PART-OF-SPEECH TAGGING WITH NEURAL NETWORKS**].

These part of speech taggers are very important in many modules of word processors and grammar checkers. Grammar checkers totally depends upon part of speech taggers. Without part of speech tagger, the grammar checker module is incomplete. Apart from that there is always a possibility for a specific word to have more than one part of speech tag [**Rule-Based Part of Speech Tagging for Pashto Language**]. POS tagging is not just assigning different tags to different words but there is a possibility of ambiguous words in any natural language.

Previously part of speech POS tagging was a very difficult task because all of the work regarding POS was done manually with too much effort and struggle [**Rule-Based Part of Speech Tagging for Pashto Language**].

**Background**

Most of the work associated to natural language processing NLP has been done in many different languages. For instance, English, German, Chinese, Arabic and Spanish [**Rule-Based Part of Speech Tagging for Pashto Language**]. As discussed in the above section that previously many approaches have been used to make part of speech (POS) taggers. In 1971 a system was developed namely TAGGIT. It was a rule-based system which was used for the initial tagging of Brown corpus. This was a successful system because it gave an accuracy of 77% In the brown corpus [**PART-OF-SPEECH TAGGING WITH NEURAL NETWORKS**].

Another idea for part of speech taggers are hybrid taggers. Hybrid taggers can be defined as the combination of both rule-based and statistical methods. An algorithm was developed from Persian language. This algorithm was a combination of some features of the rule-based taggers and probabilistic approach to tag the unknown words. This algorithm achieved a total accuracy of 97%. [**Rule Based Part of Speech Tagging of Sindhi Language**]

Talking about the hybrid systems, the combination of rule-based and statistical model another Part of Speech tagger was developed by Altunyurt for Turkish language. Along with statistical and rule-based approach two other features were also used due to which the model obtained a total of 85% accuracy [**Rule Based Part of Speech Tagging of Sindhi Language**].

Another very commonly used and important approach is rule-based approach. As it is obvious from its name that rule-based approach will definitely use some rules in part of speech taggers in order to reduce or completely remove the ambiguity. An Arabic system (part of speech tagger) was developed by Shereen. The system was comprised of a Tagset that contained 131 tags.[ [**Rule Based Part of Speech Tagging of Sindhi Language**].] Tagging was done manually and after that a new corpus was developed for Arabic language. The Tagset were applied to that newly created corpus if the system failed to identify a word or a word is ambiguous i-e having more than one part of speech tag, the rules which were developed earlier will be used to remove the ambiguity [**Rule Based Part of Speech Tagging of Sindhi Language**].

**Data Set**

The dataset we are using for our part of speech tagger contains more than 50,000 lexicons of Pashto language. The dataset is in Xml format.

**Algorithm**

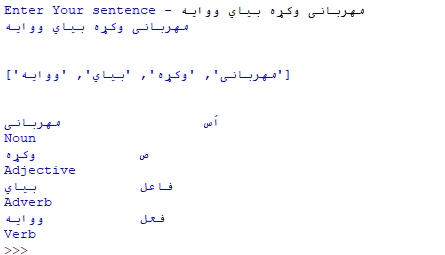
1. Take Input
2. Apply Word tokenization
3. Save each and every word in an array
4. Fetch every word from the array one by one
5. Search if the word exists in dataset
6. If yes do some preprocessing in the dataset file
7. Display the word with its associated POS as output
8. Else display this word does not exits

**Methodology**

The system is implemented in python 3.7. program takes input from the user. The input will be always a Pashto text. The input is first tokenized and each word in the text in treated as a separate entity. The tokenized words are then stored in an array. Every word is fetched from that array one by one and is checked in the existing database. If a word exists in the database then there will be some information related to that word will also exist that is gender, number, singular or plural etc. some preprocessing need to be done in the dataset file before the POS is displayed as output. So, after preprocessing if the system finds a specific word it will display that word along with its associated part of speech tag as output. Associated part of speech tag could be a noun, verb, adjective or any other POS.

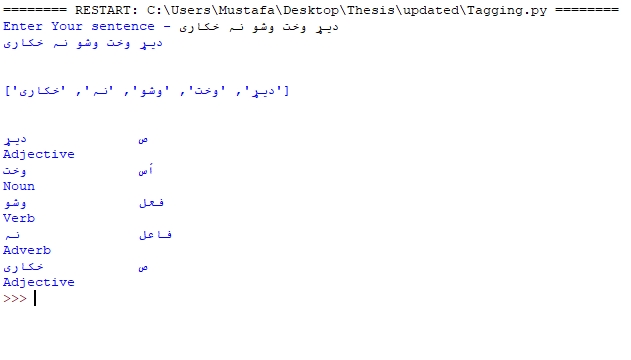
**Example figure 1**

“*Please say it again”*

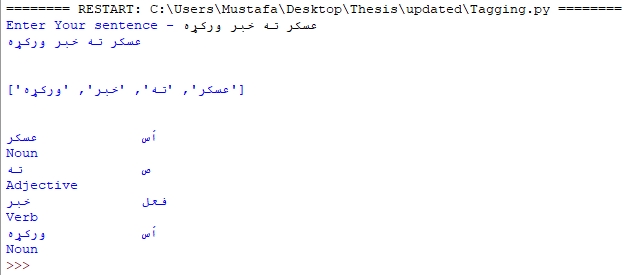


**Example Figure: 2**

*“Long time, No see”*

 ***Example Figure: 3***

*“Call the police!”*



**System’s Flow Chart:**

