International Journal of Computer Theory and Engineering, Vol. 2, No. 4, August, 2010
1793-8201

Sindhi Part of Speech Tagging System Using Wordnet

Javed Ahmed Mahar and Ghulam Qadir Memon

***Abstract*—Sindhi is highly homographic language, the text is written without diacritics in real life applications, that creates lexical and morphological ambiguity. It is a most critical problem facing Sindhi computational processing and difficult to assign correct syntactic category in the text. Lot of work has been done for diacritic restorations by using statistical and linguistics approaches, still results are not on acceptable level. Tagging the non-diacritic words can be solved using semantic knowledge. This paper describes a rule-based semantic Part of Speech (POS) tagging system that relies on a WordNet to identify the analogical relations between words in the text. The proposed approach is focused on the use of WordNet structures for the task of tagging. POS tagging is a process of assigning correct syntactic categories to each word. Tag set and word disambiguation rules are fundamental parts of any POS tagger. In this research, the tagset for Sindhi POS, word disambiguation rules, tagging and tokenization algorithms are designed and developed. Two types of lexicons are used, one for simple words and other one for disambiguated words. The corpus is collected from a comprehensive Sindhi Dictionary; the corpus is based on the most recent available vocabulary used by local people. The experiments using combination of two lexicons that show promising results and the accuracy of our proposed approach is acceptable.**

***Index Terms*—WordNet; Part of Speech; Morphology; Lexicon; Tagging Rules**

I. Introduction

Part of Speech tagging is a process of assigning syntactic categories like noun, pronoun, verb and adjective to each word in the text document. POS Tagging is an essential part of Natural Language Processing (NLP) applications such as speech recognition, text to speech, word sense disambiguation, information retrieval, semantic processing, parsing, information extraction and machine translation. Taggers can be divided as supervised or unsupervised: supervised taggers are based on pre-tagged corpora, whereas unsupervised taggers use those methods through which automatically tags are assigned to words. Furthermore, Jurafsky [3] divides taggers into three classes: (i) Rule Base Taggers: two phases are needed for these taggers. Firstly, words are searched from the lexicon and secondly, if word has assigned more than one tags then by using some syntax rule try to disambiguate it (ii) Stochastic Taggers: the probabilistic methods are used to assign a tag to the word. (iii) Transformation Based Taggers: in these taggers, assign the most probable tag to the word. After that, if wrong tag is found then by applying some rules tagger tries to change it. All these three approaches can be used with supervised as well as unsupervised taggers.

Sindhi words are polymorphemic in nature like Arabic, Urdu and Hindi. In Sindhi derivational morphology is the combination of a word stem with a grammatical morpheme usually resulting in a word of different class. For example nouns are derived from verbs and adjectives are derived from nouns and vice versa. Apart from affix changes Sindhi derivational morphology takes place by the changes of diacritic symbols. The derivational morphology example are: ٿولھ /fatness/ (noun / **اسم**), ٿلھو /fatty/ (adjective / صفت) similarly سگھارو /powerful/ (adjective / صفت), سگھ /power/ (noun / اسم) Rahman [16].

It is observed that few numbers of words are exists in languages having no fixed meaning, their meanings can be judged when these are used in context. Therefore, contextual information is necessary for understanding the word sense. As a lexical resource, WordNet offer extensive exposure of the lexicon. Knowledge of words lies not only in their meaning but also in the context in which they occur. Linking words to appropriate senses provide the desired conceptual information [18]. WordNet includes an impressive number of semantic relations like hypernymy/ hyponymy, meronymy/ holonymy, antonymy, entailment, etc. The semantic analysis of Sindhi text proves that different types of relations exist among words like noun/verb, noun/adjective, adverb/verb, etc. In this research, we have developed 368 WordNet structures especially for non-diacritic words. It is tried to collect group of analogical words of one specific word. Many NLP applications rely on syntactic information of text, but similar syntactic patterns may introduce different semantic interpretations. Likewise, similar meanings can be syntactically realized in many different ways. Therefore, Rule-based semantic POS tagging attempts to solve this problem. Our developed POS tagger relies on analogical knowledge about words drawn from WordNet for retrieving the semantic information.

Rule-based semantic systems are difficult to construct because enormous efforts are needed. We have designed two lexicons: a lexicon of Sindhi words (SWL) having no ambiguity and a WordNet lexicon (WNL) of Sindhi words having ambiguity due to absence of diacritics. Linguistics rules are also needed for appropriate and correct tagging. Sindhi POS tagging is a difficult and much more ambiguous task due to the absence of diacritic symbols. In this regard, supervised corpora were used for this tagging system. For disambiguation, rules were applied grammatically and semantically. If disambiguation was not made ever, then heuristic rules were applied. These rules are formed from

Javed Ahmed Mahar is with the Department of Computer Science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan (email:mahar.javed@gmail.com).

Ghulam Qadir Memon is with FEST, HIIT, Hamdard University, Karachi, Sindh, Pakistan (email: gqmemon@yahoo.com).

the literature and linguists. In this paper, the approach of WordNet used for the development of Sindhi POS tagger is presented. The tagset contains 67 tags & compiled. The corpus was manually tagged because it is fact that the manual construction of WordNet is reliable and gives acceptable level results, as for as linguistic soundness and accuracy is concerned it is time consuming and expensive.

A. Diacritization Problem in Sindhi

The Sindhi writing system, is based on Persian Arabic Script. Sindhi adds its own modifications in order to symbolize the many sounds not found in Arabic or Persian. The phonological systems of Sindhi in most respects resemble that of other Indo-Aryan languages. Sindhi has a very rich sound inventory. It has 43 distinctive consonant phonemes and 10 vowels. The phoneme is usually pronounced as an alveolar tap, though occasionally reminiscent of a trill with two or more contacts. There are three short vowels [a, i, u] and five long vowels [aa, ii, uu, e, o]. Short vowels are not part of the alphabet but are written as vowel diacritics above or under a consonant to give it its desired sound. There are also two diphthongs [ai, au], but these are infrequent and many dialects pronounce these the same as [e, o].

Diacritization defines the sense of words. Sindhi script consists of two classes of symbols: letters and diacritics. Letters are always written whereas text is written without diacritics in real life applications. Diacritics are extremely useful for readability and understanding, the absence creates lexical and morphological ambiguity. The absence of the diacritics in Sindhi text is one of the most critical problems facing computational processing. In this research, term diacritization is used due to the fact that the missing symbols do not represent only the vowels but also represent some other symbols. As diacritics are not written mostly in many manuscripts, therefore, words are ambiguous to understand the correct meaning of the word. For this, tagging is difficult for such kind of words. For example, a word consisting of two letters like (كن), i.e., 'k' and 'n', has ambiguity because it is written without diacritics. Nine types of homonymy words are made/ available in Sindhi dictionaries. Syntactically and semantically all these nine words are different from each other due to the placement of diacritic symbols. Therefore, it is difficult to assign correct syntactic categories to these words. Different orthographic formations of كن are shown in Table 1.

TABLE 1. Orthographic formations of كن

Sindhi Word	Transliteration	POS	Meaning
كَنُ	Kanu	Noun	Ear
كَنَ	Kana	Noun	Ears
كِنُ	Kinu	Noun	Dirt
كِنَ	Kina	Noun	Dirt's
كُنُ	Kunu	Noun	Whirlpool
كُنَ	Kuna	Noun	Whirlpools
كِن	Kini	Pronoun	Many
كَن	Kani	Pronoun	
كُن	Kuni	Noun	Cesspool

II. Related Work

In the last few years, several approaches have been developed for POS taggers of English and other Western languages. Similar work has also been done for South Asian languages. These languages have several POS taggers that use different mechanisms.

A rule-based approach uses some rules and a lexicon to resolve the tag ambiguity. Shereen [9] developed Arabic part of speech tagger and compiled a tagset containing 131 tags. After manual tagging, a lexicon of Arabic corpus is developed. When system fail to recognize word from the lexicon morphological rules are applied for correct tagging. Rabbi et al [14] proposed tagset for Pashto and developed a rule-based part of speech tagging system.

Hidden Markov Models are statistical methods which choose the tag sequence which maximizes the product of lexical probability and the contextual probability. Cutting et al [1] presents HMM based tagger for English. Some resource requirements are needed for robustness and accuracy. They achieved more than 96% accuracy and describe three applications like phrase recognition, word sense disambiguation and grammatical function assignment for tagging. Anwar et al [6] presents preliminary achievements of HMM for solving the problem of Urdu part of speech tagging system. This HMM is derived from the combination of lexical and transition probabilities. Some smoothing techniques with HMM are also used to resolve the data sparseness problem. Bar-Haim et al [7] developed a segmenter and a tagger for Hebrew based on HMM. EI Hadj etal [8] presents an Arabic part of speech tagger that uses an HMM model with the combination of morphological analysis to represent the linguistic structure of the sentence and they obtained an accuracy of 96%. Azimizadeh [11] implements the Persian part of speech tagger based on HMM. The average accuracy of this tagger is 95.11%.

A combination of both statistical and rule-based methods has also been used to develop hybrid taggers. Shamsfard et al [5] presents tagging algorithm in which combines the features of probabilistic and rule-based taggers to tag Persian unknown words. Approximately 97% accuracy is achieved. Ma et al [2] describes hybrid system for Thai part of speech tagging that consists of a neuro tagger and a rule-bases corrector. The accuracy of 99.1% is achieved which is higher than the HMM and rule-based approaches. Altunyurt et al [13] presents a composite part of speech tagger for Turkish in which they combine the rule-based and statistical approaches with two additional features and achieved approximately 85% accuracy.

Some other machine learning models are also used for tagging. Maximum Likelihood Estimation approach and some heuristic rules were used to improve the accuracy of Persian POS tagging Mohtarami et al [4]. Overall tagging accuracy is around 95.29%. Maximum Entropy Model with few contextual features is used by Ratnaparkhi [10] for part of speech tagging system of English and achieved accuracy of 96.6%. Habash et al [12] presents an approach to using morphological analyzer for tokenization, part of speech tagging and morphological disambiguation in Arabic and obtain an accuracy score of 98.1%. Sajjad et al [15] applied four probabilistic taggers i.e., TnT tagger, Tree tagger, RF tagger and SVM tool on Urdu language and achieved around 95% accyracy. Qiu et al [17] uses semantic definition in Hownet, bi-directional parallel analogy and

paired parallel analogy methods for automatic tagging.

From the very early days of language study, scholars have noted a complex network of relations existing between the word and its meaning. It has been observed that a single idea or sense can be expressed by multiple words, and conversely, multiple related ideas and senses can be expressed by a single word. The former is known as synonymy, while the latter is called polysemy [19].

Recently, many efforts have been initiated for building WordNet. Darja [20] presented an approach to automatically generate WordNet synset from the JRC-Acquis parallel corpus. The corpus was done manually. He also presented experiment in which synset for Slovene WordNet induced automatically from several multilingual resources. The evaluation of the results has shown that the method works best for nouns. Elkateb et al. [21] [18] introduces a lexical resource in the shape of WordNet for Modern Standard Arabic. The work is focus on WordNet representation of senses, word meanings and linked the AWN with SUMO where concepts are defined with machine interpretable semantics in first order logic. Shi [22] integrates three different lexical resources: FrameNet, VerbNet and WordNet into unified richer knowledge base for robust semantic parsing because each lexicon may handle different kind of problem.

III. Sindhi Tagset

The language tagset represents parts of speech and consist on syntactic classes. According to contextual and morphological structure, natural languages are different from each other. Therefore, it is necessary to have a tagset for the Sindhi language before developing a part of speech tagger. Two types of approaches are used for part of speech tagging. The contextual information is used for rule-based approach and manually assigns a part of speech tag to a word. According to part of speech, many words are ambiguous in Sindhi language. For example, a word اڌ "half" can be a noun or an adjective. It depends on the context of the sentence. If the sentence is "I ate half meal" مون اڌ مانيُ کاڌي in this sentence word اڌ is used as an adjective, on the other hand, if the sentence is "I ate half and left half" مون اڌ کاڌو ۽ اڌ بچايو in this sentence word اڌ is used as a noun. In this type of situations SPOS tagger uses context to assign parts of speech to words. Therefore, both approaches are used to reduce the size of the lexicon.

Granularity of the tagset is a major problem during the designing of tagset. The general part of speech and morpho-syntactic approaches are used for Sindhi tagset. The grammar of Sindhi has been standardized for centuries. In the Sindhi grammar books, linguistics discussed eight main parts of speech for Sindhi language. The grammarians described all features of the language but unfortunately they mixed semantic, syntactic and morphological features when they categorized parts of speech. The general parts of speech are noun, pronoun, verb, adjective, adverb, preposition, conjunction and interjection. Six parts of speech are further divided into sub-categories. Nouns are divided into noun (NN), proper noun (PNN, diminutive noun (DN), number noun(NBN), circumstance noun (CTN), connected noun (CNN), abstract noun (AN), common noun (CN), participle noun (PTN), gerund noun (GN), material noun (MN), concrete nount (CCN), sound noun (SN), present noun (PSN), object noun (OBN), ovation noun (OTN), subject noun (SJN), inflection noun (INN), collective noun (CLN), compound noun (CPN). Pronouns are divided into pronoun (PN), personal pronoun (PPN), first person pronoun (FPP), second person pronoun (SPP), third person pronoun (TPP), demonstrative pronoun (DP), reflexive pronoun (RPN), interrogative pronoun (IP), relative pronoun (RP), answer relative noun (ARP), indefinite pronoun (IDP). Verbs are divided into verb (VB), intransitive verb (IV), passive verb (PV), auxiliary verb (AV). Adjectives are divided into adjective (ADJ), proper adjective (PADJ), possessive adjective (PSADJ), positive degree (PD), comparative degree (CD), superlative degree (SD). Adverbs are divided into adverb (ADV), adverb of time (AT), adverb of place (AP), adverb of manner (AM), adverb of negative (ANG), adverb of number (AON), adverb of quantity (AQ), adverb of reason (AR), interrogative adverb (IA), relative adverb (RA). Prepositions are divided into preposition (PP), compound preposition (CP), phrase preposition (PPS). Tag for conjunction and interjection are (CC) and (IJ) respectively. Few morpho-syntactic tags are also used for training corpus. Numerals are divided into cardinal (CA), ordinal (OR), fractional (FR), multiplicative (MUL). For measuring units (MU) tag is assigned to words. Titles are frequently used in Sindhi. Therefore, two tags (PRT) for pre-title and (POT) for post title are assigned to words. Two types of markers are used in the text: sentence marker (SM) and phrase marker (PM). English style date and time is used in the Sindhi text. Therefore, (DATE) and (TIME) tags are used for this purpose.

IV. Analogical WordNet

The lexical resources are always required for supervised rule-based taggers and Word Sense Disambiguation (WSD) systems. WordNet has played important role for WSD especially in terms of semantic relations. Lexemes in WordNet have been organized around the semantic relations of synonymy, homonymy, antonymy, hyponymy/ troponymy and meronymy [23]. In wordnet database, words are divided through part of speech and put in order to hierarchy of nodes. Each node represents a concept and words denoting the same concept are grouped into a synset with a unique id. Linking words to appropriate senses provides the desired conceptual information [21].

Sindhi is highly homographic in that non-diacritized words those look similar orthographically, but contain different meanings according to diacritization. For instance, a word سر can be written by using diacritic symbols as سِرَ /a brick/ (noun), سِرُ /the head/ (noun), سُرُ /tone/ (noun), سُرَ /tones/ (noun), سُرْ /move/ (verb/noun), سَرُ /a kind of reed/ (noun), سَرَ /a string of beads/ (noun), سَر /head, principal/ (noun), سِرْ /secret/ (noun), سِر /at, on, upon/ (adverb), each word represents different meaning and different part of speech. Due to the absence of diacritic symbols, it is impossible to tagging the correct syntactic categories on that

kind of words without the semantic knowledge.

The semantic information can be retrieved by analysis of analogical relations between the words. Consider the following set of words.

ڳاءِ، ڳائي، ڳائيان، ڳائيندو، ڳائيندي، ڳائيندا، ڳائيندڙ، ڳائيندڙون، ڳائين، موسيقي، ڳائنڊ , ڳارائنڊ , ڳارائي ,راڳ، ڳائبو، ڳائيندڙ، ڳائيندڙاسين

Each word of this list has analogical relation with words سُرُ or سُرَ. This list of words also shows the individual relations between the words and they defined with respect to a coherent chunk of common sense background information. For example, consider the sentence استاد منظور علي خان ڏاڍو سر سان ڳائيندو هيو. /Master Manzoor Ali Khan sang very melodiously/. The analogy is سر: ڳائيندو . Consider another sentence. هر فنڪار کي راڳ سر سان ڳائڻ گهرجي / Every artist should sing a lyric melodiously./ The analogical words of this sentence are راڳ : سر : ڳائڻ . It has been proved that through the process of analysis of analogical words we can easily assign most appropriate tags to words even they have not used diacritic symbols in the text. Figure 1 shows the basic formation of كني having three letters, the six derived words from this formation and the analogical words of each derived word.

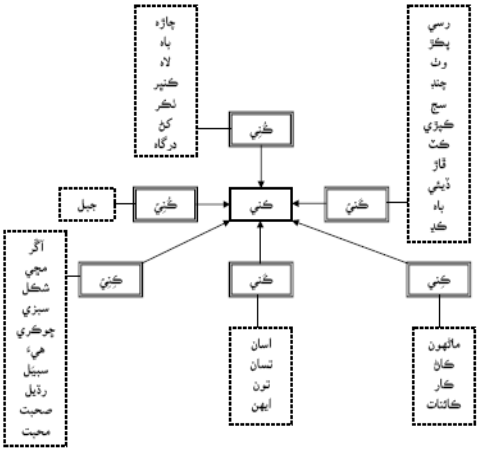

FIGURE 1. Formation of كني, derived words and analogical words

V. POS Tagging Algorithm

Take input text

1. Tokenize input text.
2. Store all words into array WORD

Select each word one by one from array WORD.

3. Search and compare selected word from array CRITICAL of words having ambiguity.
4. If word is found, assign it a separate id then
5. Select all associated id words from adjacent array HOMONUMY. Else go to step 10.
6. Select all other words from array WORD and compare them with lexicon WNL
7. If comparison is successful then select appropriate word from array HOMONYMY and go to step 15. Else go to step 10.
8. Search and compare selected word from lexicon SWL
9. If word is found one or more times, then store associated tag or tags of word into array TAGS
10. Else display "the word is not found", add this new word with corresponding tag into lexicon. Add linguistic rules for new word.
11. If one tag is stored in array TAGS, then display word with associated tag as an output.
12. Else select one or more linguistic rules and search most appropriate tag for a word by applying rules
13. Display word with associated tag as an output.

VI. Tokenizer

The tokenization is the process of segmenting input sequence of orthographic symbols. The division of input text into tokens is important for POS tagging. The Sindhi orthography is based on the concatenation of morphemes. The words are delimited through white spaces. The segmentation of words sometimes may be ambiguous in terms of part of speech tagging.

The input text is a sequence of characters which encode a sequence of words:

$$< w1,w2,w3, ...wN > \rightarrow < c1,c2,c3, ...cM >$$

The sentence is segmented by using white space because the occurrence of white space indicates the existence of a word boundary. There are various morphological problems where this approach fails. This is all about the orthographical behavior of Sindhi. For example, if the sentence is احمد منهنجو پٽ آهي. /Ahmed is my son/. There is no any ambiguity because each word easily can be segmented by using white space. Many compound words are found in the Sindhi lexicon. Compound words are those which are formed by combining two or more words. The formation of compound words is ambiguous in terms of tokenization because compound word has soft or hard space between their sub-parts. Consider the word سنڌوندي /Indus River/, we can see a soft space is present between the سنڌو /Indus/ and ندي /River/ both are also separate words in the lexicon. In Sindhi alphabet, there are two types of characters: connector and non-connector. Connector characters automatically concatenate with next character like ب، ل whereas non-connector characters are not concatenated with succeeding character like د، ڈ. The character و is also a non-connector character so that we can see a soft space. In the Sindhi text editor, system does not store this soft space in memory. Therefore, this type of space do not create problem for POS tagging. On other hand, if the word is اسلام عليكم here we explicitly put the hard space between اسلام and عليكم otherwise this word is visualize like اسلامعليكم and this is not correct formation in Sindhi language. This explicit hard space is affecting the POS tagging. For example, word صاحب قدرت is an adjective and this word is the combination of two words صاحب /Lord/ (noun) and قدرت /Nature/ (noun). If we write this word without explicit hard space then it looks like صاحبقدرت. This is not a correct formation of the word. Therefore, how to tag two tokens which make up a single morphological word.

Various solutions have been considered to solve this problem. A new tokenization scheme is proposed and implemented and the accuracy of 98.6% has been achieved. In this scheme, two variables SPACE1, SPACE2 are created. SPACE1 is used for simple hard space, which we normally put through space bar between two words. SPACE2 is used for inter space, which we put through tab key between the subparts of compound word. For this purpose, tab key is reset on single space.

Tokenization Scheme

1. Take input text
2. Create an empty list WORD1 and WORD2
3. Create variables CHARACTER, SPACE1 and SPACE2
4. Assign SPACE1=32 for space bar key and SPACE2=9 for tab key
5. Set starting point of the sentence
6. Move forward and examine each character
7. Assign each character to variable CHARACTER
8. If CHARACTER = SPACE1, then set word boundary at this point
    a. If WORD2 is empty then append all characters into WORD1 from the starting point to the boundary point and go to step 6.
    b. Else append all characters into WORD2
    c. Select two words from the WORD2 and concatenate previous word & SPACE2 & next word and append into WORD1.
    d. Delete all characters from WORD2.
9. Else If CHARACTER = SPACE2, then append all characters into WORD2 and go to step 6.
10. Else If CHARACTER = Sentence Marker then Exit, otherwise go to step 6.

VII. Lexicons

A lexicon is a declarative data structure which contains word entries and language rules. Most of the POS tagging systems are widely used lexicon. We studied linguistic facts carefully from Sindhi grammar books and observed that a word may have more than one part of speech. For example, a word افسوس /regret/ may be حرف ندا (interjection), اسم (noun) or صفت (adjective). The context of sentence is important for examining the part of speech for a word. These types of words are stored in the lexicon according to the occurrence of parts of speech of a word in Sindhi language. In the traditional language lexicon, meaning and form of a word is stored. For POS tagging, the information of words and their corresponding part of speech are required. Therefore, two columns are defined in the lexicon data base, one for words entry and second one for possible part of speech of that word. A lexicon named SWL is developed having entries of 26366 words. The initial corpus is gathered from a comprehensive Sindhi dictionary.

Two types of approaches were proposed by researchers for understanding the genuine sense of word. One is knowledge based approach and other one is corpus based approach. In this research, corpus based approach is preferred because we try to receive required information from the corpus. For this purpose three types of arrays are used for handling non-diacritic words having ambiguity. (i) CRITICAL: array used for critical words like سر , كن , كني , قسم , جوء having entries of 46 different formations of words (ii) HOMONYMY: array used for all possible words which are stored in array CRITICAL. Total entries of this array are 213. (iii) WNL: array is used for all analogical words of all words which are stored in array HOMONYMY. The total numbers of analogical words in WordNet lexicon are 1885.

A. Lexicon Ambiguity

The orthographical representation of many words may identical but they have distinct part of speech and unrelated meanings, i.e. homonymy. As in English the word *bank* where the first sense refers to a river side and the second to a financial institution. Similarly, in Sindhi a word مان /I/ refers different senses according to the placement in the sentence. For example, in the sentence مان اچان ٿو /I am coming/ the word مان /I/ refers to first person pronoun. Whereas in the sentence کيلن ۽ انبن مان کيلا ڪڻ /among bananas and mangoes take bananas/ the word مان /among/ refers to preposition. Polysemy occurs in these cases where a word has multiple senses. Ambiguity and polysemy need important attention because they affect the organization of word meaning and its part of speech. This problem can be solved with the representation of lexical knowledge for the clarity of senses. The Sindhi synsets is developed with hypernym relations to form a semantic hierarchy. Most of the work has done manually.

VIII. Tagging Rules

As we have discussed in section 4, more than one part of speech can be assigned to a single word. Many words are exist in Sindhi lexicon those are candidates for multiple tags. The ambiguity of words in term of part of speech can be removed by using grammatical rules. For this purpose, set of 186 disambiguation rules are used for SPOS tagging system. Some disambiguation rules with examples are described below.

Example 1.

هڪڙو ماڻهو آيو.

“A man came”

هڪڙو آيو ۽ ٻيو ويو.

“Somebody came and the other one went”

The word هڪڙو may be adjective as in the first sentence or noun in the second sentence. The disambiguation rule for this word is:

Given Input “هڪڙو “

If (+1 NN)

Then eliminate NN tag

Else if (+1 VB)

Then eliminate ADJ tag

Example 2.

هو آيو.

“He came”

هو ماڻهو آيو.

“The man came”

The word هو is pronoun in the first sentence and adjective in the second sentence. The disambiguation rule for this word is:

Given Input “هو “

If (+1 VB)

Then eliminate ADJ tag

Else If (+1 NN)

Then eliminate PNN tag

In Sindhi language, one rule can be applicable on more than one word. The rule for example 2 is also used for words کو، کهڙو، پنهنجو .

Example 3.

مان هڪ اڻ واقف نوجوان سان مليو هيس.

"I had met a stranger young man"

The words اڻ واقف is a compound noun but if this word appears before noun then the part of speech will change into adjective.

Given Input "اڻ واقف "

If (+1 NN)

Then eliminate CPN tag

Else eliminate ADJ tag

Example 4.

پلو گهوڙو ڊوڙي ٿو.

"An energetic horse runs well"

گهوڙو پلو ڊوڙي ٿو.

"The horse runs smart"

The word پلو is an adjective in the first sentence and adverb in the second sentence. The disambiguation rule for this word is:

Given Input "پلو "

If (+1 NN)

Then eliminate ADV tag

Else if (+1 VB)

Then eliminate ADJ tag

Example 5.

غلط ماڻهو کي سنو نه چئو.

"Don't say noble to wrong mam"

ماڻهو غلط ٿي سگهي ٿو.

"Man can be wrong"

The word غلط is an adjective in the first sentence and adverb in the second sentence. The disambiguation rule for this word is:

Given Input "غلط "

If (+1 NN)

Then eliminate ADV tag

Else if (+1 VB)

Then eliminate ADJ tag

IX. Experiments And Results

For testing, two kinds of corpora were used for Sindhi POS tagger: (i) the words having no ambiguity (ii) the words having ambiguity due the absence of diacritics. The efficiency of Sindhi POS tagging system is measured for Sindhi TTS system. In this regard, Sindhi corpora were compiled and used for testing. The training corpus is gathered from "The Comprehensive Sindhi Dictionary" because all linguists of Sindhi language are agreed that most Sindhi words are found in this dictionary. The corpora were manually tagged with our own tagset. For testing, the words have been taken from the daily Kawish news papers. From this news paper 1500 sentences were selected for testing.

The Sindhi POS tagger was tested on our own developed lexicons SWL and WNL. The lexicon SWL contains 26366 tagged words, among them the frequency of noun is 15091, pronoun is 137, verb is 4656, adjective is 5328, adverb is 979, preposition is 98, conjunction 18 and interjection is 59. The frequency of words is graphical represented in figure 2. The lexicon WNL contains 1885 analogical words. The analogical words are collected during the manual analysis of Sindhi sentences. The system searched the analogy between words in the sentence. If analogical words are not sufficient for getting semantic information in terms of non- diacritic critical word identification then the analysis process of tagging system is switched from sentence to paragraph. This switching process may continue for analysis and understanding to examine the correct sense of critical word. Only those sentences were considered having critical words. The lexicon for critical words CRITICAL and the lexicon for homographic words HOMONYMY were also developed. The combinations of these three lexicons were used for handling non-diacritic words having ambiguity. During training and testing data, we have classified the words according to the part of speech and calculated the accuracy of each type of words. The cumulative accuracy of 96.28% was achieved without use of WordNet. The evaluated results are shown in Table 2. The cumulative accuracy of 97.14% was achieved with WordNet. The evaluated results are shown in Table 3. The clear accuracy difference of four syntactic categories i.e., noun, verb, adjective and adverb are found during experiments. This difference of achieved accuracy is graphically shown in figure 3. Only 46 different words formations were used for the development of WordNet. When more different formation will be included in the lexicon WNL then the accuracy of tagger may increase.

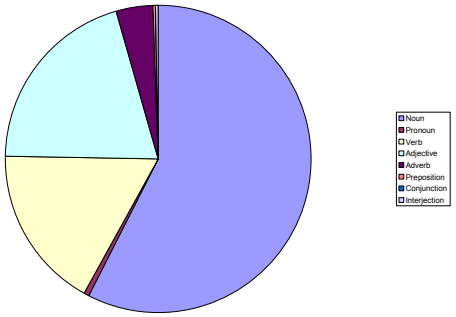

FIGURE 2. Graphical representation of words frequency

TABLE 2. Calculated accuracy of words frequency without WordNet

Part of Speech	Training Corpus	Testing Corpus	Unknown Words	Accuracy %
Noun	15091	3414	176	94.845
Pronoun	137	34	2	94.11
Verb	4656	1766	11	99.37
Adjective	5328	1112	28	97.48
Adverb	979	429	13	96.97
Prepositions	98	16	2	87.5
Conjunction	18	5	0	100
Interjection	59	7	0	100
Total	**26366**	**6783**	**232**	**96.28**

TABLE 3. Calculated accuracy of words frequency with WordNet

Part of Speech	Training Corpus	Testing Corpus	Unknown Words	Accuracy %
Noun	15091	3414	71	97.92
Pronoun	137	34	2	94.11
Verb	4656	1766	4	99.77
Adjective	5328	1112	8	99.28
Adverb	979	429	6	98.6
Prepositions	98	16	2	87.5
Conjunction	18	5	0	100
Interjection	59	7	0	100
Total	**26366**	**6783**	**93**	**97.14**

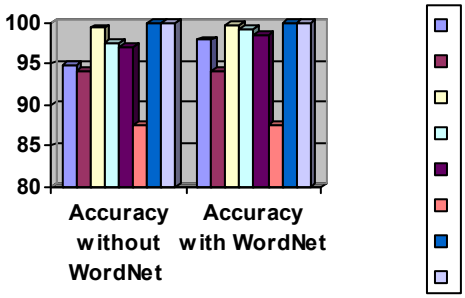

FIGURE 3. Graphical representation of accuracy differences

The GUI of Sindhi POS is classified into three main windows: (1) Sindhi text editor (2) Output window (3) Unknown words window. The system read Sindhi text from right to left; when terminator symbol is found then system compare each word of sentence with lexicon entries. If word is found then three types of information will display: (i) word (ii) part of speech in English (iii) part of speech in Sindhi. If word is found having multiple tags then system search most appropriate rule for the word. By applying rule, system selects most appropriate tagged word according to the context of sentence. If word is not found in the lexicon, then system marks it as unknown word and displays it in the unknown word window. As for as non-diacritic critical words are concerned, tagger automatically tag words as noun. During the study of Sindhi language, it is found that the 60% to 70% of words having noun syntactic category. Consider the first sentence of the example, words سِندُ and سَندُ are individuals having different meaning but both contains noun category. The output example of Sindhi POS tagger is shown below:

سنڌ جي تاريخ پنج هزار سال پراڻي آهي. سنڌي بولي دنيا جي قديم بولين مان هڪ بولي آهي. هن ڏرتيءَ تي ڪيترائي مشهور بزرگ پيدا ٿيا آهن.

سنڌ	Noun	اسم
جي	Relative Pronoun	ضمير موصول
تاريخ	Noun	اسم
پنج	Number Noun	اسم عدد
هزار	Number Noun	اسم عدد
سال	Noun	اسم
پراڻي	Adjective	صفت
آهي	Verb	فعل
.	Terminator	جملي جي ختم ٿيڻ جي نشاني

سنڌي	Proper Adjective	صفت خاص
بولي	Noun	اسم
دنيا	Noun	اسم
جي	Relative Pronoun	ضمير موصول
قديم	Adjective	صفت
بولين	Noun	اسم
مان	Preposition	حرف جر
هڪ	Number Noun	اسم عدد
بولي	Noun	اسم
آهي	Verb	فعل
.	Terminator	جملي جي ختم ٿيڻ جي نشاني

هن	Pronoun	ضمير
ڏرتيءَ	Noun	اسم
تي	Preposition	حرف جر
ڪيترائي	Adverb	ظرف
مشهور	Adjective	صفت
بزرگ	Noun	اسم
پيدا	Verb	فعل
ٿيا	Helping Verb	مددي فعل
آهن	Verb	فعل
.	Terminator	جملي جي ختم ٿيڻ جي نشاني

X. Conclusion

Part of Speech tagging is an important component of Natural Language Processing applications. Before the development of Sindhi POS, the literature of Sindhi language was reviewed in terms of morphological and syntactical perspectives. In this paper, the tagset, lexicon and word disambiguation rules were discussed and presented. Part of speech tagging and tokenization algorithms were developed and implemented. The orthography of Sindhi language is complex due to the absence of diacritic symbols. Therefore, supervised approach was used for the development of Sindhi part of speech tagging system. For handling non-diacritic words having ambiguity, the approach of WordNet were used. Initially, the accuracy of 96.28% was achieved but after applying WordNet approach the accuracy of tagger was increased up to 97.14%. If more critical words would be inserted in WordNet, the accuracy may increase up to 98%. During the experiments, it has been observed that the accuracy of tagger was low when we tested poetry text and sentences of future tense. Similarly, when we tested sentences of simple present and past tenses then the accuracy was very high. The validation test was performed by linguists of Sindhi. They are fully satisfied with the outputs of Sindhi POS. The future work of this study will be for statistical approaches of SPOS and then compare results with rule based approach.

References

[1] Cutting, D., Kupiec, J., Pederson, J., Sibun, P., (1992) "*A Practical Part-of-Speech Tagger*". In "Proceedings of the Third Conference on Applied Natural Language Processing", pp. 133-140.

[2] Ma, Q., Murata, M., Uchimoto, K., Isahara, H., (2000), "Hybride Neuro and Rule-Based Part of Speech Taggers", International Conference on Computation Linguistics", pp. 509-515.

[3] Jurafsky, D., J.H. Martin, (2000), "SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistic and Speech Recognition", Prentice-Hall.

[4] Mohtarami, M., Amiri, H., Oroumchain, F., Rahgozar, M., (2008), "Using Heuristic Rules to Improve Persian Part of Speech Tagging Accuracy", 6th Int. Conference on Informatics and Systems (INFOS2008), pp. 34-38.

[5] Shamsfard, M., Fadaee, H., (2008), "A Hybrid Morphology-Based POS Tagger for Persian", Proceeding of the 6th International Language Sources and Evaluation, pp. 3453-3460.

[6] Anwar, W., Wang, X., Luli, Wang, X., (2007), "Hidden Markov Model Based Part of Speech Tagger for Urdu", Information Technology Journal 6(8), pp. 1190-1198.

[7] Bar-Haim, R., Sima'an, K., Winter, Y., (2005), "Choosing an Optimal Architecture for Segmentation and POS Tagging of Modern Hebrew", In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages.

[8] El Hadj Y. O. M, Al-Sughayeir, I. A., Al-Ansari, A. M., (2009), "Arabic Part of Speech Tagging Using the Sentence Structure", Proceedings of the 2nd International Conference on Arabic Language Resources and Tools".

[9] Shereen Khoja, (2001), "APT: Arabic Part of Speech Tagger". In NAACL Student Workshop.

[10] Ratnaparkhi, (1996), "A Maximum Entropy Model for Part of Speech Tagging", In Proc. of the Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania.
[11] Azimizadeh, A., Arab, M. M., Quchani, S. R., (2008), "Persian Part of Speech Tagger Based on Hidden Markov Model", 9th JADT.
[12] Habash, N., Rambow. O., (2005),"Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop". Proceedings of the 43rd Annual Meeting of the ACL, pp. 573-580.
[13] [13] Altunyurt, L., Orhan, Z., Gungor, T., (2007), "Towards Combining Rule-Based and Statistical Part of Speech Tagging in Agglutinative Languages", Computer Engineering Vol 1. pp. 66-69.
[14] [14] Rabbi, I., Khan, M.A., Ali, R., (2009), "Rule-Based Part of Speech Tagging for Pashto Language", Proceedings of the Conference on Language & Technology", pp. 82-87.
[15] Sajjad, H., Schmid, H., (2009), "Tagging Urdu Text with Parts of Speech: A Tagger Comparison", Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 692-700.
[16] Rahman, M. U., (2009), "Sindhi Morphology and Noun Inflections", Proceedings of the Conference on Language & Technology", pp. 74-81.
[17] Qiu, L., Zhang, X., Mao, L., (2009), "Building A Dictionary on Constituent Structure of Chines Compounds", Proceedings of the Conference on Natural Language Processing and Knowledge Engineering, pp. 9-16.
[18] Elkateb, S., et al. (2006), "Arabic WordNet and the Challenges of Arabic", in proceedings of Arabic NLP/MT conference.
[19] Desh, N. S., (2004), "Corpus-based Study of Lexical Polysemy in Bangla for Application in Language Technology", in the proceedings of the SIMPLE, pp. 70-74.
[20] Darja Fiser, (2007), "A Multilingual Approach to Building Slovene WordNet", workshop on common Natural Language Processing Paradigm.
[21] Elkateb, S., et al. (2006), "Building a WordNet for Arabic", in the proceedings of the fifth international conference on language resources and rvaluation.
[22] Shi, L., Mihalcea, R., (2005), "Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing", computational linguistics and intelligent text processing, pp. 100-111.
[23] Langone, H., Haskell, B.R., Miller, G. A., (2004), "Annotating WordNet", in proceedings of the workshop on frontiers in corpus annotation.
[24] Jurafsky, D., J.H. Martin, (2000), "SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistic and Speech Recognition", Prentice-Hall.