Rule-Based Part of Speech Tagging for Pashto Language

Ihsan Rabbi, Mohammad Abid Khan and Rahman Ali
Department of Computer Science, University of Peshawar, Pakistan
ihsanrabbi@gmail.com, abid_khan1961@yahoo.com, rahmanali.scholar@gmail.com

Abstract

In natural language processing, part-of-speech tagging plays a vital role. It is a significant pre-requisite for putting a human language on the engineering track. Before developing a part-of-speech tagger, a tagset is required for that language. This paper is about the first ever rule based part-of-speech tagging system for Pashto language and a tagset that helps in the development of a Parser for the said language [8]. A very simple architecture is applied that gives reasonably good accuracy.

1. Introduction

Part of speech tagging (POS tagging) has a vital role in different fields of natural language processing including machine translation. Part-of-speech tagging is defined as the process of assigning to each word in a running text a label which indicates the status of that word within some system of categorising the words of that language according to their morphological and/or syntactic properties (frequently known as a “tagset”) [2].

Earlier POS tagging was done manually with much effort, but nowadays automatic POS tagging mechanism is becoming more common. POS tagging is not just giving tags to words but in any natural language many words are ambiguous. For example, English word ‘Eye’ in the following sentences gives different parts-of-speech.

Miss Hina has beautiful eyes.
Special Forces eyed the whole area.

In the first sentence, the word ‘eye’ is a noun whereas in the second sentence it is a verb. Also, the parts of speech are not just the eight or nine such as noun, pronoun, verb and adverb. Their sub-categorization is also done such as a noun may be singular or plural, common or proper and a pronoun may be a personal pronoun, a demonstrative pronoun or a possessive pronoun. Before developing a POS tagger, it is necessary to have a tagset for the language.

The next section includes some related techniques of POS tagging in other languages that may help in Pashto as well. Subsequently, there is a proposed tagset and architecture for POS tagging. Lastly, some real data is taken and is tested on the given POS tagger and the statistics is summarized accordingly.

2. Related work in other languages

Most of the work related to natural language processing has been done in languages such as English, German, Chinese and Arabic. These languages have several part-of-speech taggers that use different mechanisms. For instance, English language has tagsets developed using rules based, statistics, transformational based and artificial neural network based [13] [15]. Erwin Marsi et al have developed POS tagging for Arabic language [6]. Daniel Tianhang Hu has designed POS tagging for Chinese [7]. Urdu has its own Part of Speech tagger that was developed by Andrew Hardie [2].

3. Proposed tagset

There are different tagsets available in different languages. For English, many people define their own tagsets. Common examples of English tagset are used for Brown corpus, Penn Treebank tagset, C5 tagset and C7 tagset [13]. For Urdu, Andrew Hardie developed a tagset [2] and for Arabic language Shereen Khoja, Roger Garside and Gerry Knowles developed a tagset [5].

Every natural language has some differences with other languages. Therefore, there is a need for a separate tagset of Pashto language. The proposed tagset has 54 tags where there are 8 tags for nouns, 12 tags for pronouns, 7 tags for verbs, 17 for punctuations, two for adpositions, and 1 for each adjective, adverb, conjunction, interjection, number, foreign word, abbreviation and symbol.

A noun may be common or proper, may be singular or plural and direct or oblique. Similarly, pronouns may be personal, demonstrative, possessive, indefinite or reciprocal. Other parts of speech are also divided into their subcategories. The full tagset is given in the appendix with proper Pashto example.

4. Proposed architecture

Before discussing the architecture, a very simple algorithm for POS tagging is defined as below which shows the basic structure of the proposed architecture.

4.1. Algorithm for POS tagging

1 Take input text
2 Tokenize the input text
3 Search for the tokens in lexicon
4 If not found, mark those tokens
5 Tag all tokens using rules if multiple tags exist
6 Get the tagged output text
7 Extract those marked tokens from the tagged output
8 Insert those new words in lexicon.
9 Add rule for that new word.

The proposed architecture for POS tagging is given below.

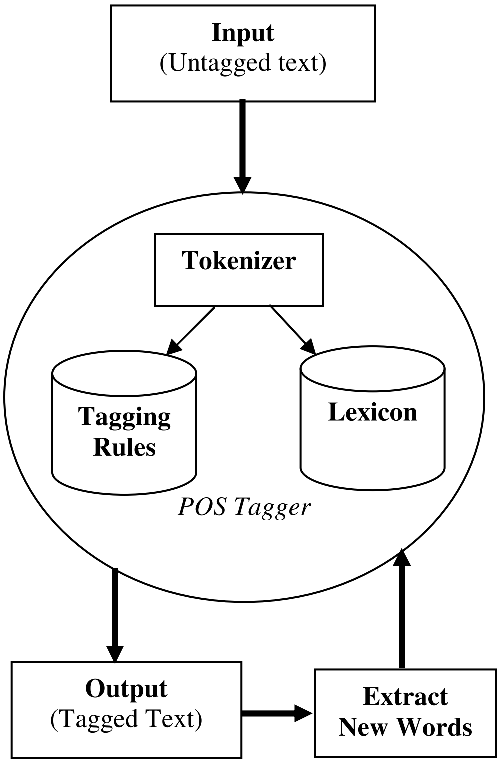

Figure 1: Part of speech tagging architecture

The functionality of each module in architecture is now explained one by one.

4.2. Tokenizer

In POS tagging, the first entity is the division of input text into different tokens. In the above architecture, the functionality of the tokenizer is to identify tokens in the source text. As no proper corpus is available for Pashto language so the formation of tokens is a bit difficult. The white space is mainly used for the tokenization of input text. However, in Pashto text many words are ambiguous in their formation because some single words have spaces between their subparts. For example,

کراچۍ یو ډیر ښکلے او عجیبه ښار دے -

[kračəy] [yew] [ḍəyr] [ṣ̌kulay] [aw] [ςǰibə] [ṣ̌ar] [dəy].

[Karachi] [one] [very] [beautiful] [and] [wonderful] [city] [to-be (copula)].

“Karachi is one of the most beautiful and wonderful cities.”

خویندو وروڼو یو بل سره لوبی کولې -

[xwəyndo] [woroṇo] [yewbal] [srə] [lobe] [kawləy].

[sisters] [brothers] [each other] [with] [play] [to-do].

“Sisters and brothers were playing with each other.”

In the first sentence, the word ‘یو’ means one (that is a numeral) while in the second sentence the word ‘یو’ is combined with ‘بل’ to form the reciprocal pronoun (‘یو بل’) meaning each other. The correct tokenization of such words requires rules. A simple rule for the above problem is as:

“if ‘یو’ + 1 is ‘بل’ then combine ‘یو’ and ‘بل’ else consider ‘یو’ as individual token”

Similarly, rules are necessary for the identification of e.g. ‘–‘ (end of sentence in Pashto text), open quotation, closed quotation, open parenthesis and closed parenthesis. All these punctuations are not tokenized by space because no space is given between them and other word. Rules have been developed for such type of identification. A simple rule for such tokenization is:

“if there exists a punctuation with other word then make the word as separate token and the punctuation as well”

Some other rules are mentioned in other subsections.

4.3. Lexicon

The lexicon used here is the simplest one [13]. Lexicon contains words and the corresponding part-of-speech tags taken from the tagset. Initially, manual tagging is performed for a small set of Pashto text. Manual tagging consists of all the words that belong to close classes tags such as prepositions, conjunctions, interjections and pronouns. Lexicon will grow when real text will be used that is taken from sources such as the newspapers, books and Internet on this POS tagger and get the words tagged. Then those words that are new to the lexicon are extracted from the tagged text and inserted into the lexicon.

4.4. Extraction of new words

A word that is currently not available in the lexicon may be extracted from the output text. The mechanism is to first tag those words that are in the lexicon whereas those words that are not in the lexicon are given a specific mark. Afterwards, those new words are tagged using rules. When the extraction is done, it is displayed and checked manually. Corrections are done to those tags that are incorrect and are then entered into lexicon.

4.5. Tagging rules

Most words in any language take multiple tags. Assigning a specific tag to a word requires rules. In Pashto text, there exist words that are candidates for multiple tags. Consider the following example in this regard.

د یو بل نه به مو د خپلی خپلی صوبی او علاقی متعلق معلومات کول -

[da] [yewbal] [nə] [bə] [mo] [da] [xpali] [xpali] [ṣobəy] [aw] [ҫlaqəy] [motaҫliq] [maҫlomat] [kawal].

[from] [each other] [to] [future-particle] [ours-clitic] [from] [own] [own] [province] [and] [area] [about] [information] [take].

“We got information from each other about each other’s province and area.”

Consider another example.

دغسی زمونږ یو بل دوست خبره کوي -

[daγasi] [zamonğ] [yewbal] [dost] [xabarə] [kawi].

[similarly] [ours] [another] [friend] [talking] [do].

“Another of our friend has the same notion.”

The word “یو بل ” may be Reciprocal Pronoun (as in the first example) or Adjective (as in the second example).

Now the rule for disambiguation is as

Given input: “یو بل ”
If (+1 is Noun)
Then assign Adjective tag
Else assign Reciprocal Pronoun tag

Similarly, consider other examples.

هغه مسکے شو او لاړ په خپله مخه -

[haγə] [muskəy] [šo] [aw] [laṛ] [pə] [xpalə] [maxə]

[he] [laugh] [to-become] [and] [went] [to] [own] [way]

“He smiled and went on his way.”

Consider a further example:

اوس به درله هغه جامی راوړم چی ناویانی ئی اغوندي -

[aos] [bə] [darlə] [haγə] [ǰaməy] [raoṛam] [če] [naoiane] [ye] [aγondi].

[now] [future-particle] [you (directional-pronoun)] [that] [cloths] [bring] [that] [bride] [her (clitic)] [wear].

“Now, I shall bring those clothes for you that are worn by brides.”

The word’هغه ’ may be personal pronoun (as in the first example) or demonstrative pronoun (as in second example).

The rule of disambiguation is:

Given input: “هغه”
If (+1 is Noun)
Then assign Demonstrative Pronoun tag
Else assign Personal Pronoun tag

On the same token, consider the following examples.

په دی موسم کښی به خلق نمر ته ډير زيات خوشحاليدل -

[pə] [dəy] [mosam] [kṣ̌e] [bə] [xalaq] [namar] [tə] [ḍəyr] [ziat] [xošḥaləydal].

[on] [that] [season] [in] [future-particle] [people] [sun] [to] [very] [much] [pleased].

“People used to be very happy to the sun in this season.”

هغه مور ډېر منع کولو چي بچی دا ته په کومه روان ئې دا ښهٔ لار نهٔ ده
-

[haγə] [mor] [ḍəyr] [mançe] [kawalo] [če] [bačəy] [da] [tə] [pə] [komə] [rawan] [ye] [da] [ṣ̌ə] [lar] [nə] [də].

[he] [mother] [very] [forbid] [to-do] [that] [son] [it] [you] [on] [which] [go] [to-be] [it] [good] [way] [not] [to-be].

“He was always stopped by his mother from the way he was following as it was not the right one.”

In the same way

یه خوری ! ته ګوره کنه هغه بله ورځ ئی زما زامنو ته وئیلي دي چی
د ماماګانو سره مو د مور زمکه ده ۔

[yə] [xore] [!] [tə] [gorə] [kanə] [haγə] [balə] [wraj] [ye] [zama] [zamano] [tə] [wayali] [di] [če] [da] [mamagano] [srə] [mo] [da] [mor] [zmakə] [də].

[o] [sister] [!] [you] [see] [tick] [that] [other] [day] [him (clitic)] [mine] [sons] [to] [talk] [to-do] [that] [to] [uncles] [has] [their (clitic)] [her] [mother] [land] [to-be].

“O sister! You see that it has been said to my sons that their mother has a share in their uncles’ property.”

The word ‘ته’ may be personal pronoun or a postposition. In the first sentence, it is used as a postposition whereas in second sentence it is a personal pronoun. In the third sentence, its both forms exist.

Now the rule for disambiguation is as

Given input: “ته”
If (-1 is Noun)
Then assign Postposition tag
Else assign Personal Pronoun tag

Similarly, new words that may not be present in the lexicon can also be the input of POS tagger. So, assigning a proper tag for the new word may also be considered. We know that the word order of Pashto sentences is SOV [1]. When a word is scanned and that word is not in lexicon, it is highly likely that the given word will be a noun or verb because these are open classes. So when a new word occurs that is not in our lexicon it will be tagged according to given rules. For verb identification the rule may be as

Given input: “New word”
If (+1 is Punctuation/Auxiliary verb/Copula verb)
Then assign verb tag
Else assign noun tag

5. Test and result

Initially, a very limited lexicon and rules were present in POS tagger. So, when we performed POS tagging in our POS tagger, its accuracy was low. However, when more text is tagged and manual corrections are done for those words that are new words to lexicon, the lexicon will grow and rules will be added for those new words. After some time, the accuracy level will also be increased. When our lexicon is reached to 100,000 words and rules to 120, the accuracy became 88%. All these are summarized in the Table 1.

Table 1: Results of POS tagger

No. of Words in Lexicon	No. of rules	POS tagger Accuracy
100	10	40%
1000	40	62%
10,000	70	76%
100,000	120	88%

From the above table, it is clear that the accuracy is increased with increasing the number of words in the lexicon and the number of rules for disambiguation.

6. Conclusion and future work

To enable most of our common people to communicate with the computer and use a computer more efficiently, it is necessary to enable computers to know our language. Part of speech tagging plays a vital role in natural language processing. This paper presents a reasonably accurate POS tagger for Pashto language.

Part of Speech tagging helps in the creation process of a parser [8], therefore the future work includes the design of a parsing algorithm for Pashto language. The structure of Pashto sentences must be identified during the development of Pashto parser.

7. References

[1] F. Babrakzai, “Topics in Pashto Syntax”, Ph.D Thesis, Linguistics Department, University of Hawai’I, 1999.

[2] A Hardie, “The computational analysis of morpho-syntactic categories in Urdu”, PhD Thesis, Department of Linguistics and Modern English Language, University of Lancaster, 2003.

[3] M. A. Khan, and F. Zuhra "A Computational Approach to the Pashto Pronoun", PUTAJ Science, University of Peshawar, 2006.

[4] F. Zuhra, and M. A. Khan, "Towards the Computational Treatment of the Pashto Verb", In Proceeding of Conference on Language and Technology (CLT07), Brar Gali Summer Campus, University of Peshawar, August 7-11, 2007.

[5] S. Khoja, R. Garside, and G. Knowles, “A tagset for the morpho-syntactic tagging of Arabic”, Computing Department Lancaster University

[6] E. Marsi, A. V. Bosch, and A Soudi. “Memory based morphological analysis generation and part of speech tagging of Arabic”, Center for Computational Linguistics Rabat, Morocco.

[7] D. T. Hu, “*Development of Part of Speech Tagging and Syntactic Analysis Software for Chinese Text*”, Bachelors & Masters thesis of Engineering in Electrical Engineering and Computer Science.

[8] M. Dalrymple, “How much can part-of-speech tagging help parsing?”, Centre for Linguistics and Philology, University of Oxford, Oxford OX1 2HG UK.

[9] H. Penzl, "A Grammar of Pashto", University of Michigan, February 1954 .

[10] S. Reshteen, “Pashto Grammar”.

[11] T. Rahman, “The Pashto language and identity-formation in Pakistan”, http://www.informaworld.com/smpp/title~content=t713411866

[12] H. Tegey, and B. Robson, "A Reference Grammar of Pashto", Center for Applied Linguistics, Washington, D.C, 1996.

[13] D. Jurafsky, and J. H. Martin, “*Speech and Language Processing”, Published by Pearson Education (Singapore).*

[14] M. A. Zyar, "*Pashto Grammar*", Danish Publishing Association Qissa Khwani Bazar, Peshawar, 2003.

[15] H. Schmid, “Part-Of-Speech Tagging With Neural Networks”, Institute for Computational Linguistics, Azenbergstr.12, 70174 Stuttgart, Germany.

[16] M. J Yar, “*Gul Meena*”: A Pashto Novel published by Pashto Academy University of Peshawar, ISBN 969-418-026-0

Appendix

Tags	Description	Example
N1CD	Singular Common Direct Noun	هلک [halak] Boy
N1CO	Singular Common Oblique Noun	هلک [halak] Boy
N2CD	Plural Common Direct Noun	هلکان [halakān] Boys
N2CO	Plural Common Oblique Noun	هلاکانو [halakāno] Boys
N1PD	Singular Proper Direct Noun	--
N1PO	Singular Proper Oblique Noun	--
N2PD	Plural Proper Direct Noun	--
N2PO	Plural Proper Oblique Noun	--
PP	Personal Pronoun	زه [zə] I
PS	Possessive Pronoun	زما [zmā] Mine
PD	Demonstrative Pronoun	دغه [dáγa] This
PI	Indefinite Pronoun	هر [her] Every
PX	Reflexive Pronoun	خپل ځان [xpəl jān] Self
PR	Reciprocal Pronoun	یوبل [yew bəl] Each Other
PN	Intensive Pronoun	پخپله [paxplə] Oneself
PG	Interrogative Pronoun	څوک [cúk] Who
PL	Relative Pronoun	کوم یو

		[kúm yaw] Which One
PT	Directional Pronoun	را [rā] I
PC	Clitic Pronoun	مې [me] I
PB	Distributive Pronoun	هریو [her yew] Each One
VS	Present verb	خورم [xoram] Eat
VT	Past Verb	وخوړه [woxṛə] Ate
VI	Infinitive Verb	ګرځیدل [garjedal] Walking
PF	Future Particle	به [bə] Indicate Future
VA	Auxiliary Verb	کول [kawal] To Do
VSC	Present Copula Verb	یم [yəm] To Be
VTC	Past Copula Verb	وم [wom] To Be
Pre	Preposition	د [da] From
Pos	Postposition	کښې [kše] In
Num	Numeral	یو [yew] One
Int	Interjection	هائي [haəy] For Sorrow
Con	Conjunction	او [aw] And
Adv	Adverb	نژدي [niždəy]
Adj	Adjectives	ښکولي [školəy] Beautiful
FW	Foreign Word	--
Abb	Abbreviation	--
Sym	Symbol	--
.	Full Stop	۔
,	Comma	،
?	Question Mark	؟
!	Exclamation Mark	!
:	Colon	:
;	Semi-Colon	;
“	Open Quotation Mark	”
”	Closed Quotation Mark	“
(	Open Parenthesis	)
)	Closed Parenthesis	(
[	Open Square Bracket	]
{	Open Curly Bracket	}
}	Close Curly Bracket	{
]	Closed Square Bracket	[
"	Neutral Quotation	"
‘	Open Single Quotation Mark	’
’	Closed Single Quotation Mark	‘