ALGORITHM FOR THE CLITICIZATION OF CONTEXT DEPENDENT PRONOUNS IN PASHTO LANGUAGE

Azizud Din[1],[2] , Bali Ranaivo-Malancon[2], and Alvin W. Yeo[2]
[1]Al Jouf University, KSA, aziz621@gmail.com
[2]UNIMAS, Malaysia, mbranaivo@fit.unimas.my, alvin@fit.unimas.my

ABSTRACT. The replacement of strong pronouns with counterpart weak pronouns (Citics) is an important task in the translation of Pashto into other languages by computer before anaphora resolution can take place. Repetition of proper nouns and common nouns is not a good way in a language. Instead of it, pronouns are used in most languages but in Pashto language weak pronouns are mostly used in context dependent text. Especially in poetry often clitics are used instead of strong pronouns. The presence of clitics and strong pronouns at the same time in Pashto language complicates anaphora resolution. Replacing strong pronouns with clitics makes the text very simple and efficient. In Pashto, some pronouns are context free and some are context dependent. Context free pronouns can be replaced with clitics using simple rules that encompass a single sentence in which the pronoun itself occurs. Replacement of context dependent strong pronouns with corresponding clitics involves syntactic agreement across single or multiple clauses. In this paper, an algorithm is presented for the cliticization of context dependent strong pronouns which backtracks to previous adjacent clause(s) to replace context dependent strong pronoun with clitics using syntactic constraints.

Keywords: Clitics, Morpheme, Cliticization, Weak Anaphoric Reduction Process-WARP

INTRODUCTION

A clitic is a morpheme that has syntactic characteristics of a word but shows evidence of being phonologically bound to another word. It syntactically functions as a free morpheme but phonetically appears as a bounded morpheme. In Pashto, a word and a clitic attached to this word are pronounced as a single word, while in written text Clitics are often written as separate words. Syntactically, a clitic, together with the word to which it is bound, functions above the clause level. Clitics attach only phonetically to the first, last or to the only word in a phrase, clause or whichever part of speech the word belongs to.

Morphologically, Pashto clitics are neither independent words nor affixes. They follow the host word to which they are associated. Generally, their placement in the phrase or a sentence is based on the syntactic rules of the language. Linguistically, clitics forces NP (noun phrase) reduction process, also termed as WARP (Weak Anaphoric Reduction Process) (Tagey, H. 1977). At the discourse level, Pashto clitics are used for emphasizing focus on either subject or object. Clitics occur in various positions in sentences, except in the start. Normally, a clitic occurs in the second position of the clause, i.e. second position from the right of the clause (Babrakzai, F. 1999). Table 1 gives a complete list of context dependents demonstrative pronouns used in Pashto language.

Table 1. Pashto dependent strong pronouns

Pashto Context dependents pronouns	Gloss	Type
دهغه	Da agha	Possessive Demonstrative with preposition
دهغي	Da aaghi	Possessive Demonstrative with preposition
دهغوئ	Da aghoi	Possessive Demonstrative with preposition
په هغه	Pa agha	Demonstrative with preposition
په هغي	Pa aghi	Demonstrative with preposition
په هغوئ	Pa aghoi	Demonstrative with preposition
هغه	Agha	Demonstrative pronoun
هغي	Aghi	Demonstrative pronoun
هغوئ	Aghoi	Demonstrative pronoun
ور	Wer	Oblique Pronominal

In the cliticization process, strong pronouns are replaced with semantically equivalent weak pronouns (clitics) (Din, Khan (2007). The key advantages of the cliticization process are: It reduces the domain of anaphoric devices in input text so that anaphora resolver would deal with a smaller set of pronouns, which in turn improves the performance of anaphora resolver. It helps in the translation of text into other syntactically related languages.

The replacement of pronouns with clitics in a clause alters the topicalization i.e. the Pragmatic function 'focus' may shift from subject to object and vice versa, which should be explicitly preserved before cliticization. Focus is the essential piece of information that is carried by a sentence. Focus is marked in all languages by intonation prominence (focal stress), but in many languages it is indicated by word order and/or special particles or clitics as in Pashto. Focus preservation can be done by marking the entity in a clause explicitly as topicalized, before replacing the strong pronouns (Kroeger, 2004).

CONTEXT DEPENDENT PRONOUNS

A context dependent pronouns refers to a previously mentioned constituent (normally, previously adjacent clause), and fills the position of a noun phrase in a clause. Mostly, هغه (agha), هغي (aghi), هغوئ (aghoi), هغهده(da agha), ده هغي (da aghi), ده هغوئ (da aghoi)), هغهپه (pa agha), په هغي (pa aghi), په هغوئ (pa aghoi) etc. occur in Pashto text similar to anaphoric devices because they are syntactically linked to the subject or object of the previous clause. The replacement rules for these strong pronouns have to take context into account. The following are the example sentences containing these strong pronouns. Here, the symbol # marks clause boundaries.

Example 1a. (With strong pronoun)

#کله چه سليم ځی# نو ځه د هغه سره ځم#

نو	ځی	سليم	چه	کله
[no]	[zi:]	[sΛl i:m]	[chI]	[kΛlə]
Then	Go	Saleem		When

ځم	سره	د هغه	ځه
[zΛm]	[sΛrə]	[dəghə]	[zə]
Go	With	Him	I

When Saleem goes, I go with him.

Example 1b. (With Clitic)

#كله چه سليم ځئ# نو ځه ور سره ځم#

نو	ځئ	سليم	چه	كله
[no]	[zi:]	[sΛl i:m]	[chI]	[kΛlə]
Then	Go	Saleem		When

ځم	سره	ور	ځه
[zΛm]	[sΛrə]	[vΛr]	[zə]
Go	With	(clitic)	I

When Saleem goes, I go with him.

Following is the example sentence containing third persons possessive دهغي without postpositions.

Example 3a. (With strong pronoun)

سدره كور ته راننوزي،# نو مور دهغي په ژړا شئ# # څنګه چه

راننوزي	ته	كور	سدره	چه	څنګه
[rənΛnəzi:]	[tΛ]	[koor]	[sIdrə]	[chI]	[sΛngə]
Enters		House	Sdra		When

شئ	ژړا	په	د هغي	مور
[ʃhi:]	[jəΛrə]	[pə]	[dəgeI]	[moor]
Start	Cry	PostP	Her	Mother

As soon as Sidra enters the house, her mother starts weeping.

Example 3b. (With Clitic)

څنګه چه سدره كور ته راننوزئ# نو مور ي په ژړا شئ#

راننوزي	ته	كور	سدره	چه	څنګه
[rənΛnəzi:]	[tΛ]	[koorr]	[sIdrə]	[chI]	[sΛngə]
Enters		House	Sdra		When

شئ	ژړا	په	ي	مور
[ʃhi:]	[jəΛrə]	[pə]	[ji:]	[moor]
Start	Cry		(clitic)	Mother

As soon as Sidra enters the house, (her) mother starts weeping.

The next section describes the Cliticization of context dependent strong pronouns in detail by a computer system.

CLITICIZING CONTEXT DEPENDENT STRONG PRONOUNS

For Cliticization of Pashto text containing context dependent strong pronouns, rule based approach is used. An algorithm is developed that takes the parsed Pashto text and transformation rules as input after describing the rules of cliticization.

Following are the transformation rules for the cliticization of third person demonstrative and possessive demonstrative pronouns.

IF the POSTP in the 2nd clause is “ سره"then replace the “ د هغه،دهغوئ،د هغي , ” with “ور”

IF the POSTP in the 2nd clause is “"باندےthen replace the “ په هغه،په هغي،په هغوئ” with “ور”

IF there is no POSTP in the 2nd clause, then replace “هغه " with “ي ”

IF there is no POSTP in the 2nd clause, then replace “هغي " with “ي ”

IF there is no POSTP in the 2nd clause, then replace “هغوئ " with “ي ”

IF there is no POSTP in the 2nd clause, then replace “د هغه" with “ي ”

IF there is no POSTP in the 2nd clause, then replaceدهغي" " with “ي ”

IF there is no POSTP in the 2nd clause, then replace “دهغوئ" with “ي ”

For the strong pronouns Table 2 summarizes the replacement criteria.

Table 2. Pashto Strong pronouns’ transformation table

Preconditions			Replacement
Pronouns	Gloss	Postpositions	Clitics
د هغه	his	لاندے/سره/نه	ور
د هغي	her	لاندے/سره/نه	ور
د هغوئ	their	لاندے/سره/نه	ور
په هغه	On him	باندے	ور
په هغي	On her	باندے	ور
په هغوئ	On them	باندے	ور
هغه	Him	ته	ور
هغي	Her	ته	ور
هغوئ	their	ته	ور
د هغه	his	Nil	ي
د هغي	her	Nil	ي
د هغوئ	their	Nil	ي

The transformation rules are represented using prolog predicates for evaluation in the following table.

Table 3. Transformation Rules

Serial No.	Transformation rules
1	rule(sp(هغهدَ), rpct(ې), pos(nc)).
2	rule(sp(هغهدَ), rpct(ې), pos(nc)).
3	rule(sp(هغيدَ), rpct(ې), pos(nc)).
4	rule(sp(هغوئدَ), rpct(ې), pos(nc)).
5	rule(sp(هغه), rpct(ې), pos(nc)).
6	rule(sp(هغي), rpct(ې), pos(nc)).
7	rule(sp(هغوئ),rpct(ې), pos(nc)).
8	rule(sp(هغوئ), rpct(ې), pos(nc)).
9	rule(sp(هغهدَ), rpct(ور), postp(سره), pos(nc)).
10	rule(sp(په هغه), rpct(ور), postp(باندے), pos(nc)).

The list of abbreviations, used in table 3, is given in Table 4. Both the rules and input text will be encoded in Unicode when developing a computer program in C++.

Table 4. Abbreviations used in rule encoding

Abbreviation	Description
Ct	Clitic
Sp	Strong pronoun
Pos	Position
Postp	Postposition
ReplaceSP	Replace strong pronoun
Nc	Not change
Rep	Replacement
C	Clause

The algorithm takes the above rules and parsed Pashto text as input. The rest of this section gives algorithm listing, and detailed explanation of its working.

Algorithm: Pronoun Replacer

```
1. Tag input text.
2. Parse input text and mark  Syntactic entities.
3. Divide complex and compound   sentences into clauses.
4. FOR EACH clause Ci in the text
    BEGIN
            FOR EACH pronoun SPj in Ci
            CALL ReplaceSP( Ci, SPj)
    END
5. END.
```

Sub Module: ReplaceSP(Ci, SP)

```
FOR EACH Rule Rj in RuleSet
BEGIN
  IF (Rj.SP = SP) THEN
  BEGIN
    IF all conditions in Rj are true  for Ci AND Ci-1 THEN
      Delete SP from Ci.
      Place Rj.Rep at Position Rj.Pos
      RETURN
  END
END.
```

The main algorithm is responsible for reading, parsing, clause division, and detection of strong pronouns in the text. It starts with the reading of Pashto text in Unicode. The step-1 and 2 tag and mark each word in the text for its grammatical category. Step-3 divides complex and compound clauses are into simple clauses. After parsing and clause division, the algorithm sets a counter variable named i for processing all the clauses in the text. At each iteration of the loop a test is made to find out, if the clause Ci contains a strong pronoun SPj.

Here, j shows the strong pronoun number. For each strong pronoun SPj the algorithm calls a subprogram ReplaceSP(Ci, SPj) which is responsible for replacing the strong pronoun SPj in clause Ci of input text. When all of the clauses have been processed, the algorithm stops.

The strong pronoun replacement subprogram ReplaceSP(Ci,SP) takes two parameters, i.e. a clause and a strong pronoun. The first parameter is the Ci, which is the clause in which the strong pronoun has to be replaced. The second parameter SP is the strong pronoun which has been found in the clause Ci and needs to be replaced. At the start of the replacement process, the algorithm set a counter variable j to 0 for iterating through the RuleSet.

The counter variable j is used for indexing into a rule table (i.e table-3) designed for replacing strong pronouns. The algorithm iterates through the rule table using j as index. At each jth row of the rule table, the strong pronoun in at RuleSet[j]. SP is matched with the strong pronoun SP in Ci. If a match occurs the algorithm applies preconditions from RuleSet at jth row to the clause CL, to determine if replacement at the jth row of the rule table can be applied to the clause Ci.

If all the conditions are true in the jth row; the transformation at the jth row is applied to clause Ci. The strong pronoun SP is replaced by a clitic given in RuleSet[j].rep. The subprogram ReplaceSP stops after the replacement of the strong pronoun.

The text data contains the parsed clauses. Table 3 of rules and table 4 of abbreviations are based on these small clauses. Some of the few tested clauses are:

1a. clause(txt(کله چه سلیم ځی نو)). 1b. clause(txt(ځه ده هغه سره
ځم)).
2a. clause(txt(نو کله چه سدره راغله)). 2b. clause(txt(خور هغي
ووهله)).
The program produces the following output for the above clauses.
1ab. کله چه سلیم ځی نو ځهور سره ځم 2ab. کله چه سدره راغله نو خور
ي ووهله

EVALUATION

A corpus of 50 different sentences was evaluated after tagging and parsing. 49 sentences were correctly cliticized by the proposed algorithm. The one sentence was not successfully cliticized because of rule application ambiguity, which resulted in the problematic situation where more than one rule could be applied at the same time to cliticize a sentence containing pronoun. Manual evaluation of the algorithm showed that the algorithm did not alter the semantic structure of the input text, only focus on subject or object shifted. More over the cliticized text was found to be suitable for anaphora resolution.

CONCLUSION

Replacement of context dependent strong pronouns with corresponding clitics involves syntactic agreement across single or multiple clauses. The proposed algorithm achieves 98% accuracy in cliticizing context dependent pronouns in Pashto. The algorithm is linear time and based on a compact set of hand-crafted rules. The cliticized sentences can be efficiently used in anaphora resolution.

REFERENCES

Babrakzai, F. 1999. Topics in Pashto Syntax. *Ph.D. Dissertation*, University of Hawai'i at Manoa.

Din, Khan (2007), Syntax Based De- Cliticization of Pashto text for Better Machine Translation. *The proceedings of Conference on Language and Technology* (CLT07) at Bara Gali campus, University of Peshawar (August 7- 11, 2007) Page no.1.

Kroeger, (2004). *Analyzing Syntax A lexical functional Approach*. London: the press syndicate of the University of Cambridge, page-136.

Tagey, Habibullah. (1977) The Grammar of Clitics: Evidence from Pashto and Other Languages. *PhD Dissertation* University of Illinois.