Clause Boundary Identification using Classifier and Clause Markers in Urdu Language

Daraksha Parveen, Ratna Sanyal, and Afreen Ansari

***Abstract*—This paper presents the identification of clause boundary for the Urdu language. We have used Conditional Random Field as the classification method and the clause markers. The clause markers play the role to detect the type of sub-ordinate clause, which is with or within the main clause. If there is any misclassification after testing with different sentences then more rules are identified to get high recall and precision. Obtained results show that this approach efficiently determines the type of sub-ordinate clause and its boundary.**

***Index terms*—Clause marker, conditional random field.**

I. Introduction

CLAUSE boundary identification is a useful technique for various Natural Language Processing (NLP) applications. This is a method of specifying the beginning and ending of main and subordinate clause. Clauses are structural unit which have verbs with its arguments, adjuncts etc. There are 8 types of subordinate clause: *Complementizer, Relative Participle, Relative, Temporal, Manner, Causality, Condition* and *Nominal*. First three types of clauses are more syntactic while remaining five clauses are more semantic in nature.

Numerous techniques are used to recognize clause boundaries for different languages where some are Rule based [Harris 1997; Vilson 1998] and others are Statistical approaches using machine learning techniques [Vijay and Sobha, 2008]. A rule based clause boundary system has been proposed as preprocessing tool [Harris 1997] for bilingual alignment parallel text. In another pioneering work, a rule based system has been used which reduces clauses to noun, adjective or an adverb [Vilson 1998]. Identification of clauses for English language has been performed in an earlier research [Sang and Dejean, 2001]. A hybrid approach for clause boundary identification uses Conditional Random Fields (CRF) and rules, error pattern analyzer used to correct the false boundaries [Vijay and Sobha, 2008]. The clause identification for Tamil language shows 92.06% and 87.89% for precision and recall respectively, which in turns give the F-measure as 89.04%. The clause boundary identification has also been done for Bengali Language [Ghosh et. al. 2010]. CRF based statistical techniques are used to identify the type of clauses. The clause identification system gives the precision as 73%.

A basic clause identification system has been developed [Ejerhed 1988] for improving American Telephone & Telegraph (AT&T) text to speech system. This was used in English/Portuguese machine translation system. Clause spitting is also needed for the text to speech, which can be done by using conditional random fields' technique [Nguyen et.al. 2007]. In Korean language, analysis of dependency relation among clauses is very critical part. Kernel method [Kim et. al. 2007] is used to detect the clause boundaries. In Japanese language, there is no distinct boundary information to detect clauses; ambiguity can be minimized using rule based system [Fujisaki et.al. 1990].

In our present work, a hybrid approach is proposed that uses both techniques i.e. rule based and machine learning to build an identifier for different clause boundaries of Urdu language. We have applied the Conditional Random Fields (CRF). We have categorized the different types of sub ordinate clauses on the basis of clause markers. The POS tagger and Chunker [Pradeep et. al. 2007] are used to prepare the parts of speech and chunked tagged data as the inputs, where linguistic rules are taken as features. To the best of our knowledge, no work on identification of clauses for Urdu language is reported.

Henceforth presented details are divided into the following sections. We have given the introduction with related work in section 1. The methodology with clause markers, Clause Boundary Annotation Convention, Preprocessing, classification with features and rules are discussed in section 2. In the example sets, the Urdu sentences are translated in English for the easiness of the readers who are not familiar in Urdu. The algorithms for different phases are given in section 3. Section 4 shows the result of clause identification for Urdu language using this algorithm. Section 5 comprises the conclusion and finally reference section is included at the end.

II. Methodology

We have prepared the corpus for Urdu language. POS tagging and chunking are the preprocessing steps which have been done manually here, so contain a great accuracy. The POS and chunked tagged corpus has been considered as input data. Initially machine learning approach is applied, within which linguistic rules are used. Through this, clause boundary

Manuscript received November 2, 2010. Manuscript accepted for publication January 12, 2011.

The authors are with Indian Institute of Information Technology - Allahabad, India (e-mail: daraksha.parveen3022@gmail.com; rsanyal@iiita.ac.in; afreen.aa@gmail.com).

is recognized from input Urdu corpus. Now, if there is any misclassification, correction is done through additional linguistic rules. The work flow of identification of clauses is shown in Fig. 1.

We have used the CRF techniques as modeling in the learning phase and inference in the classification. This is a sequential classification technique which is taking care of many correlated features like in Maximum-entropy and a variety of other linear classifiers including winnow, AdaBoost, and support-vector machines [Sha et.al. 2003]. CRF gives more beneficial results than HMMs on a part-of-speech tagging task [Lafferty et.al. 2003]. Hidden Markov Model (HMM) needs to enumerate all possible observation sequences. This is not practical to represent multiple interacting features or long-range dependencies of the observations. Also it has very strict independence assumptions on the observations [Kelly et.al. 2009].

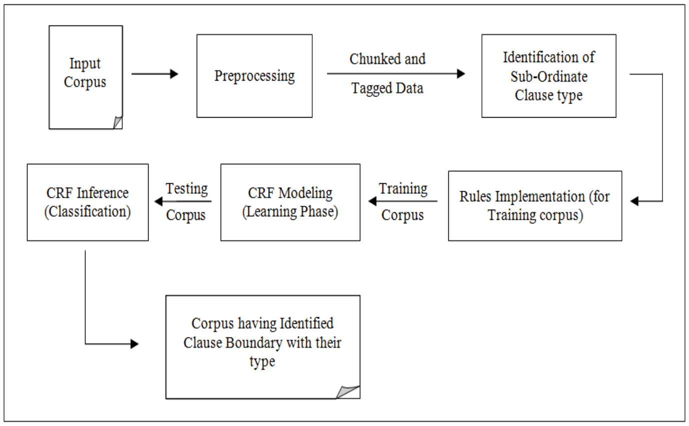
Input Corpus
Preprocessing
Chunked and Tagged Data
Identification of Sub-Ordinate Clause type
CRF Inference (Classification)
Testing Corpus
CRF Modeling (Learning Phase)
Training Corpus
Rules Implementation (for Training corpus)
Corpus having Identified Clause Boundary with their type

Fig. 1. Work flow for the identification clauses.

CRF uses the conditional probability *P (label sequence* **y** | *observation sequence* **x***)* rather than the joint probability *P*(**y**, **x**) as in case of HMM. It specifies the probability of possible label sequences **y** for a given observation sequence **x**. CRF allows arbitrary, non-independent features on **x** while HMM does not. Probability of transitions between labels may depend on past and future observations.

The shallow parsing uses special kind of CRF technique where all the nodes in the graph form a linear chain. In this type of graph, the set of cliques C (a graph in which every two subset of vertices are connected to each other) is just the set of all cliques of size 1 (i.e. the nodes) and the set of all cliques of size 2 (the edges). This technique has two phases for clause boundary identification:

1. *Learning*: Given a sample set X containing features $\{x_1, \dots, x_N\}$ along with the set of values for hidden labels Y i.e. clause boundaries $\{y_1, \dots, y_N\}$, learn the best possible potential functions.
2. *Inference*: For a given word there is some new observable x, find the most likely clause boundary y* for x, i.e. compute (exactly or approximately):

$$y^* = \arg \underbrace{max}_{y} P(y|x) \qquad (1)$$

For this, an undirected and acyclic graph formed which contains the set of nodes $\{x_i\} \cup Y$ $(\forall x_i \in X)$, adopts the properties by Markov, is called conditional random fields (CRFs). Clause Boundary Detection is a shallow parsing technique so, CRF is used for this.

A. Clause Markers

Clause markers are words or a group of words, like *now* and *well* in English, which helps in making the relation between the sentences. They are also used in combining two Urdu sentences as shown below.

```
(i) [Ram ghar aya] aur [khana kha kar so gaya]
    [Ram came home] and [fell asleep after eating
dinner]
```

As discussed earlier there are 9 types of subordinate clause. There are clause markers corresponding to these subordinate clauses which are more syntactic. For relative participle clauses, clause markers are *jo vo* (وہ جو), *jisne usne* (اُسنے جسنے), *jinhe* (جنھیں) etc. In the relative participle clause *vo* will always occur with *jo* as correlation [Butt et.al. 2007] as shown below

```
(ii)[Jo ladka kal aya tha][vo ram hai]
    [That boy [who came yesterday] is Ram]

(iii)[vo ladka [ jo kal aya tha] ram hai]
     [Ram is the guy][who came yesterday]
```

Similarly, conditional clause markers, complementizer clause markers, and relative clause markers are shown below as boldface letter sequentially.

(iv)

$$\underbrace{[\textbf{Agar}\ \text{kal tum nahi aye}]\,[\ \textbf{to}\ \text{mein khane ke baad so jaunga}]}_{\text{Conditional clause}}$$

[If u do not come tomorrow][then I'll sleep after eating dinner]

(v) $\underbrace{[\text{Ram ne kal kaha tha}][\ \textbf{ki}\ \text{vo ghar ja raha hai}]}_{\text{Complementizer clause}}$

[Ram said yesterday] [that he is going home]

(vi) $\underbrace{[\text{Main [khana kha kar] so gaya}]}_{\text{Temporal clause}}$

[I fell asleep after dinner]

(vii) $\underbrace{[\textbf{jisse}\ \text{mujhe kaam hai}]\,[\text{vo kahan hai}]}_{\text{Relative clause}}$

[Where is that person] [With whom I have a work]

B. *Clause Boundary Annotation Convention*

For main clause, clause boundary annotations are shown below where the symbols CL, B, M, and E are clause, beginning, main, and ending respectively.

```
CL = B_M
CL = E_M
```

For subordinate clause, clause boundary annotations are shown below where symbol 'Sub' indicates sub-ordinate clause.

```
CL = B_Sub_Type
CL = E_Sub_Type
```

Annotations for sub-ordinate clause types are shown below.

```
RELP: For Relative Participle
COMP: For Complementizer
COND: For Conditional
TMPR: For Temporal
CAUS: For Causal
RELC: For Relative
NOML: For Nominal
MANR: For Manner
```

C. *Preprocessing*

In the preprocessing stage, at first tagger is applied on the tokenized corpus to get tagged data and then chunker is applied to obtain chunked and tagged data (see Fig. 9). Further processing will be done on these tagged and chunked data. Sentence Boundaries are not given in the preprocessed data.

POS and Chunked tagged data are shown in Table I. There are three columns where first column comprises of tokens, second of tags for corresponding token and third contains chunking information. Here 'B' corresponds the beginning of phrase and 'I' to the words which are in a phrase.

TABLE I
POS AND CHUNKED TAGGED DATA

Tokens	Tags	Chunking
اَنڈین	PRPN	B-NP
پیک	NN	I-NP
مشہور	ADJ	O
ہے	VAUX	B-VP
جہاں	CC	B-CCP
پہاڑوں	NN	B-NP
پر	PSP	O
لمبے	ADJ	O
وقت	NN	B-NP
تک	PSP	O
گھومنے	VB	B-VP
کے	PSP	O
ماہرین	ADJ	O
چڑھائی	ADJ	O
چڑھ	VB	B-VP
کر	VB	I-VP
دلی	ADJ	O
مراد	NN	B-NP
پوری	ADJ	O
کرتے	VB	B-VP
ہیں	VAUX	I-VP
۔	SYM	O

D. *Classification*

Sequence labeling classification technique is applied in the clause boundary identification. Clause Identification has been done by using linguistic rules which do not depends upon sentence boundaries. Classification technique requires features, training data set and testing data set. As discussed in sec. 2.1, classification has two phases, learning and inference. In learning phase, modeling takes place by taking training dataset as an input while in inference phase; classification of test data set takes place with the help of model obtained from learning phase.

1) *Features*

In this CRF technique linguistic rules are used as features for which different length of windows, comprises of words, are formed that depend on these linguistic rules. For example, in case of relative clause identification in Urdu language, clause beginning and ending are identified via rule1 and rule2 respectively.

RULE_1:

If the current word is any relative clause marker and next word is any of the POS tags verb, pronoun, adjective, noun then the next word is marked as beginning of clause boundary as shown below

```
Position 0: Relative clause marker
Position 1: Verb or Adjective or noun or pronoun
Then 0 should be marked as beginning of
subordinate clause of type relative.
```

Where position 0 indicates the current word and position 1 is the next word.

RULE_2:

If the current word is any verb auxiliary and next word is any symbol then current word is end of corresponding subordinate clause boundary as shown below

```
Position 0: Verb phrase or Verb auxiliary
Position 1: any symbol or phrase
Then 0 should be marked as end of above
subordinate clause.
```

2) *Handling Misclassification*

There is a chance of misclassification in the clause boundary ending. If there is any misclassification then correction is done through linguistic rule, which means priority is given higher to the linguistic rules.

III. ALGORITHM FOR DIFFERENT PHASES

A. *Preparation of the Training Corpus*

Step 1: First check whether a word W coming is a clause marker or not. If it is, then detect which type of clause it is.

Step 2: Implement those rules (defined as in sec.2.4.1) which is related to above type of clause which is detected in step 1. Then through these rules find the clause beginning and ending of that clause.

B. CRF Modeling (Learning Phase)

Step 1: Parse the prepared training corpus and assign $f_1, f_2, f_3, \ldots, f_m$ to those words which follows rule 1, rule 2, and rule 3… respectively.

Step 2: Make a matrix T of size M×N where,

M = no. of features ($f_1, f_2, f_3, \ldots, f_m$)

N = no. of classes (Clause beginning, Clause ending, not boundary)

Matrix is made by parsing the corpus in which,

T_{ij}=1, if a word follow rule i and belong to class j

T_{ij}=0, if not so

In this matrix we go on incrementing every time in T_{ij}, if another word follows the same.

C. CRF Testing

Step 1: Make a matrix J of size M×1 for each word where,

M = no. of features.

J_{i1} = 1, if a word follows rule i

J_{i1} = 0, if it does not follow

Step 2: Find matrix C of size 1×N

$$C_{1\times N}= J^T_{M\times 1} \times M_{M\times N} \tag{2}$$

Step 3: Assign that class to a word which has a maximum value in matrix C.

IV. Results and Discussion

The system is tested upon a corpus which consists of Urdu language dataset. The dataset comprises of different types of subordinate clause which is POS tagged and chunked. Results are shown in Table II which contains the information of clause boundary beginning and ending where B-SUB indicates the beginning of sub-ordinate clause while E-SUB is for ending of sub-ordinate clause. We have obtained the result using clause markers through which we can easily detect the type of subordinate clause. Evaluation of our system's performance is done by calculating the precision and recall as shown in Table III.

TABLE II
Output showing clause boundary beginning and ending

Tokens	Tags	Chunking
اس	PRN	B-NP
وقت	NN	I-NP
ادہیز	ADV	O
تـر	ADV	O
باشـــــندگان	ADJ	O
تـر	ADV	O
غـلام	NN	B-NP
تھــے	VAUX	B-VP
جو	CC	B-CCP <Cl=B-SUB-RELP>
باغـات	NN	B-NP
ںیم	PSP	O
کـام	NN	B-NP
کـــرتے	VB	B-VP
تھــے	VAUX	I-VP <Cl=E-SUB-RELP>
.	SYM	O

Table III shows the comparison between different ratios of corpus taken for training and testing purpose. In the corpus (developed for this work only), there are 139 different sentences with POS and Chunked tagged related to tourism domain. It is a 3-fold cross Validation represented by set-1, set-2 and set-3.

TABLE III
Comparison of Different Ratios of Training and Testing Corpus

Training-Testing	Precision (%)	Recall (%)
<u>90%-10%</u>		
Set – 1	89.2	90.0
Set – 2	88.6	89.5
Set – 3	87.5	88.9
Average value	88.4	89.5
Standard Deviation	0.705	0.451
<u>80%-20%</u>		
Set – 1	85.2	86.7
Set – 2	86.0	87.1
Set – 3	85.5	87.5
Average value	85.6	87.1
Standard Deviation	0.331	0.327
<u>70%-30%</u>		
Set – 1	82.3	84.0
Set – 2	82.6	83.9
Set – 3	82.0	84.1
Average value	82.3	84.0
Standard Deviation	0.245	0.082

Our system works very efficiently on the similar sentences shown below

Relative sub-ordinate clause

(i)[پیرس نے اس علاقہ کی طرف کوئی توجہ نہیں دی]
-[جسسے حالت بد سے بدتر ہوتی گئی]

[Paris did not pay any attention to that area][due to which the condition get worsened.]

Relative Participle sub-ordinate clause

(ii)[اس وقت زیادہ تر باشندگان غلام تھے] [جو
باغات میں کام کرتے تھے]

[That time mostly peoples were slaves] [who worked in gardens.]

Complementizer clause

(iii)[کورٹِس کے حق میں ایک بات اور ہو گئی][کہ
یہ محض اتفاق تھ ا]

[Court was in favor of something] [that was just a coincidence.]

The problem for detecting the clause ending is coming for the following types of sentences.

Relative Participle sub-ordinate clause

(i) [جب برٹش ٹاسک فورس جو جھگڑا کے اوائل
میں ہی روانہ کر دی گئی تھی] ، [تب متعدد مواقع
پر بات چیت اور بیچ بچاؤ ناکامیاب رہے]

[When the British Task Force early in the dispute had been dispatched], [then they discuss various opportunities and failed to get rescue]

Temporal sub-ordinate clause

(ii)[ںیم] [اکھان کھا کر] سو یاگ]

[I fell asleep after dinner]

Manner Sub-ordinate clause

[میں ورجش کروں گا [جیسا کہ مجھے سکھایا گیا] ہے]
(iii)

[I'll do the exercises [as I've been taught]]

After analyzing the above sentences, we have found that the sentences where the distance of clause beginning and ending is significantly large, our system is unable to detect the clause ending correctly as shown above in first sentence. Here, big braces show the actual clause beginning and ending whereas our system is unable to detect the clause ending. For those sentences which are semantic in nature, it is difficult for our system to detect clause ending and beginning as shown above in second, third and fourth sentences.

V. Conclusion

In this paper Conditional Random Fields are used for classification of clause boundary beginning and ending and also detecting the type of subordinate clause. Here, linguistic rules are given higher priority, hence misclassification is corrected via these rules. Limitation with CRFs is that it is highly dependent on linguistic rules. Missing of these rules may lead to wrongly classified data. An improvement can be achieved in the proposed clause boundary identifier by including more sophisticated linguistic rules, clause markers for different subordinate clauses and also for those clauses which are embedded in the main clause. For future work, clause boundaries detection can be done on those sentences where distance between clause beginning and ending is significantly large and also where the sub-ordinate clauses in the sentences are semantic in nature. More linguistic rules are being identified and will be apply. This work is in progress.

Acknowledgement

The authors gratefully acknowledge the support from Indian Institute of Information Technology Allahabad and Technology Development in Indian Languages, funded by MCIT (Govt. of India).

References

[1] M. Butt, T.H. King, and S. Roth, “Urdu correlatives: theoretical and implementational issues,” in *Proceedings of the LFG07 Conference*, CSLI publication, 2007, pp. 107-127.

[2] E. Ejerhed, “Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods,” in *Proceedings of the 2nd Conference on Applied Natural Language Processing*, Austin Texas, 1988, pp. 219-227.

[3] H. Fujisaki, K. Hirose, H. Kawai, and Y. Asano, “A System for synthesizing Japanese speech from orthographic text,” in *Proc. of International Conference on Acoustics, Speech, and Signal Processing ICASSP-90*, vol.1, 1990, pp. 617-620.

[4] A. Ghosh, A. Das, and S. Bandyopadhyay, “Clause Identification and Classification in Bengali,” *in Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP, 23rd International Conference on Computational Linguistics (COLING)*, Beijing, August 2010, pp. 17-25.

[5] V. P. Harris, “Clause Recognition in the Framework of Alignment,” Mitkov, R., Nicolov, N. (eds.) *Recent Advances in Natural Language Processing*, John Benjamins Publishing Company, Amsterdam/Philadelphia, 1997, pp. 417-425.

[6] D. Kelly, J. McDonald, and C. Markham, “Evaluation of threshold model HMMS and Conditional Random Fields for recognition of spatiotemporal gestures in sign language,” in *Proceedings of the 12th international conference Computer Vision Workshops (ICCV Workshops 2009),* 2009, pp. 490-497.

[7] S. Kim, S. Park, S. Lee, and K. Kim, “A Feature Space Expression to Analyze Dependency of Korean Clauses with a Composite Kernel,” in *Proceedings of the 6th International Conference Advanced Language Processing and Web Information Technology (ALPIT 2007)*, 2007, pp. 57-62.

[8] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, “Conditional Random Fields: Probabilistic Models For Segmenting and Labeling Sequence Data,” in *ICML '03 Proceedings of the Eighteenth International Conference on Machine Learning*, 2003, pp. 282-289.

[9] V. Nguyen, “Using Conditional Random Fields for Clause Splitting,” in *Proceedings of the Pacific Association for Computational Linguistics*, University of Melbourne Australia, 2007.

[10] V.D. Pradeep, M. Rakesh, and R. Sanyal, “HMM-based Language independent POS tagger,” in *Third Indian International conference on Artificial Intelligence IICAI 2007,* 2007.

[11] E.F.T.K Sang and D. Herve, “Introduction to CoNLL-2001 shared task: clause identification,” in Walter Daelemans and Remi Zajac (eds.) *Proceedings of Conference on Computational Natural Language (CoNLL 2001),* Toulouse, France, 2001, pp. 53-57.

[12] F. Sha and F. Pereira, “Shallow Parsing with Conditional Random Fields,” in *NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Volume 1, pp. 134-141, 2003.

[13] R.S.R. Vijay and L.D. Sobha, “Clause Boundary Identification Using Conditional Random Fields,” in *Lecture Notes in Computer Science, Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, Springer-Verlag, 2008, pp. 140-150.

[14] J.L. Vilson, “Clause Processing in Complex Sentences,” in *Proceedings of the First International Conference on Language Resource and Evaluation*, vol. 1, 1998, pp. 937-943.