Clause Boundary Identification for Malayalam Using CRF

Lakshmi, S.,Vijay Sundar Ram,R and Sobha, Lalitha Devi
AU-KBC RESEARCH CENTRE,MIT Campus of Anna University, Chrompet, Chennai, India
lakssreedhar@gmail.com,sundar@au-kbc.org,sobha@au-kbc.org

ABSTRACT

This paper presents a clause boundary identification system for Malayalam sentences using the machine learning approach CRF (Conditional Random Field).Malayalam Language is considered as a 'Left branching language' where verbs are seen at the end of the sentence. Clause boundary identification plays a vital role in many NLP applications and for Malayalam language, the clause boundary identification is not yet explored. The clause boundaries are identified here using grammatical features.

KEYWORDS : Malayalam Language, CRF, Clause boundaries, Left branching Language

1 Introduction

The goal of clause identification is to divide sentences into clauses which typically contain a subject and a predicate. In NLP clause identification is considered as a shallow semantic parsing technique which can be useful in applications like Machine Translation, parallel corpora alignment, Information Extraction and speech applications. The clause boundary identification is done with the help of CRF (Conditional Random Fields) which is a statistical machine learning technique. The clauses can be identified as Main clauses and subordinate clauses. There are 8 types of subordinate clauses namely: 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. In our approach grammatical features are taken into consideration and Noun chunks are being identified instead of Nouns to study the sentence structure.Malayalam belongs to the category of relatively free order, agglutinative and morphologically rich languages. Hence the part-of speech-tagging provides more information. For the clause boundary identification we take into account the suffixes attached to the verb. In Malayalam language, the verbs are usually seen at the end of the sentence and the noun phrases take a position to the left of the verb. The subject in a sentence has possessive, nominative or dative marking.

Numerous techniques have been in use to identify the clause boundaries for different languages. Early experiments in the clause identification such as Eva Ejerhed's basic clause identification system(Ejerhed, 1988) for text to speech system, Papageorgiou's rule-based clause boundary system a preprocessing tool for bilingual alignment parallel text(Papageorgiou, 1997) . Leffa's rule-based system reduces clauses to noun, adjective or adverb, which was used in English/Portuguese machine translation system(Leffa, 1998).There were hybrid clause boundary identifying systems which uses memory based learning and post process it using rule-based system by Orasan(Orasan, 2000). The clause identification was the shared task in CoNLL-2001(Tjong et al., 2001).Carreras did a partial parsing of sentence, which makes a global inference on a local classifier and used a dynamic programming for choosing the best decomposition of sentence to clauses(Carreras et al., 2002). Carreras did a phrase recognition using perceptrons and an online learning algorithm(Carreras et al., 2003).Georgiana did a multilingual clause splitting experiment, where he used a machine learning technique and indicators of co-ordination and subordination with verb information(Puscasu , 2004).

Here we have used conditional random fields (CRF) for clause boundary detection. CRF is an undirected graphical model, where the conditional probability of the output are maximized for a given input sequences(McCallum et al., 2003).This is proved successful for most of the sequence labeling tasks, such as shallow parsing(Sha , 2003),named entity recognition task(McCallum et al., 2003). CRF was used for clause splitting task by Vinh Van Nguyen, where they have also used linguistic information and a bottom-up dynamic algorithm for decoding to split a sentence into clauses(Nguyen et al., 2007). In another experiment the clause identification was done using a hybrid method, where CRF and linguistic rules were used and cascaded by an error analyzer(Vijay et al., 2008). In Indian languages, some are Statistical approaches using machine learning techniques for Tamil language(Vijay et al., 2009) which was 75% accurate.

Henceforth presented details are divided into the following sections. An introduction with related works was presented in section 1, section 2 describes our approach, where we give our method and the rules we have used. Section 3 is about the different evaluation and results obtained.

Section 4 comprises of Error Analysis and finally the conclusion and reference section is presented.

2 Our Approach

The clause identifier for Malayalam is built using CRF a machine learning technique. The CRF technique we have used grammatical rules as one of the major feature. The preprocessing of the sentences is done for part-of speech(pos) and chunking information and morphological information. The clause identifier has to learn the sentence structures. So here we have replaced the noun phrases in the sentence with a token np after preprocessing the sentence, retaining the morphological information of the head noun. In this Malayalam clause identifier-relative participle clause (RP), conditional clause (COND) and main clause (MCL) are identified.

2.1 Conditional Random Fields (CRF)

CRF has two phases for clause boundary identification:

1.Learning: Given a sample set X containing features $\{X_{1,\ldots\ldots,}X_N\}$ along with the set of values for hidden labels Y i.e. clause boundaries $\{Y_{1,\ldots\ldots\ldots,}Y_N\}$, 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):

$$y^{*}=\mathrm{argmax}_{y}P(y|x) \qquad (1)$$

For this, an undirected and acyclic graph formed which contains the set of nodes

$\{X_i\}$ U Y(V ε X), adopts the properties by Markov, is called conditional random fields (CRF). Clause Boundary Detection is a shallow parsing technique so, CRF is used for this. Now Let o = ($o_1,\ldots,o_T$)be some observed input data sequence, such as a sequence of words in a text document, (the values on T input nodes of the graphical model). Let S be a set of FSM states, each of which is associated with a label, l (such as PERSON). Let s =($s_{1\ldots}s_T$) be some sequence of states, (the values on T output nodes).

Linear-chain CRF thus define the conditional probability of a state sequence given as follows

$$P_\Lambda(\mathrm{s}|\mathrm{o}) = \frac{1}{Z_\mathrm{o}} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, \mathrm{o}, t)\right),$$

Where Z_0 a normalization factor over all state sequences, f_k ($s_{t-1,}s_{t,}$,o,t) − is an arbitrary feature function over its arguments, and λ_k (ranging from −∞to∞) is a learned weight for each feature function. A feature function may, for example, be defined to have value 0 in most cases, and have value 1 if and only if s_{t-1} is state #1 (which may have label OTHER), and s_t is state #2 (which may have START or END label),and the observation at position t in o is a relative pronoun or a conditional marker. Higher λ weights make their corresponding FSM transitions more likely, so the weight λ_k in the above example should be positive since the word appearing is any clause marker (such as conditional or relative clause marker) and it is likely to be the starting of a clause boundary.CRF++ available in open source (Kudo ,2005) is used in our approach.

2.2 Features

The vital part in any machine learning technique is to identify a proper feature set. We have used two types of features word level and structural level. At the word level we have considered the

- lexical word,
- its part-of-speech and
- chunk.

Words with their appropriate morphological information was considered. In the clause type identification task words play a crucial part as word carries the information of the clause type. Part-of-speech information which provides the context and definition of the words in a sentence is also added as a column.In Malayalam, due to rich inflection, the part-of-speech tagging of the words contribute more information than in English. Chunking can be considered as the first step towards full parsing. Here we have taken a window of size five. The structural level features are the grammatical rules. The first column of the input represents the word, the second column is the part-of- speech tag and the next column is based on the morph analysis. The last column is based on the grammatical rules. It is described in the section below.

2.3 Rules for identifying clauses

2.3.1 Relative Participle Clauses

The relative participle clause is identified by the relative participle verb in a sentence. The relative participle (RP) verb will occur in three tense and 'a' is the relative participle suffix in Malayalam.

Example 1:

innale	vanna	penkutti	ente	anujathi	anu
yesterday	come-past-RP	girl	my	sister	copula-present

(The girl who came yesterday is my sister.)

Here the RP embedded clause is “innale vanna penkutti”, ”the girl who came yesterday”.The RP is always followed by the NP that it modifies. The position of the embedded clause is not an issue.

Beginning of the RP clause is the first subject NP preceding the word containing the RP but excludes the subject NP.

The RP clause boundary is determined as per the 3 rules given below.

Rule 1. RP verb can be followed by a noun phrase (NP) and a postposition (PSP). The noun phrase will be inflected with case markers depending on the PSP that follows.

Example 2:

avan	thottathil	ulla	pookkaLe	patti	samsarichu.
he	garden-LOC	be-RP	flower-PL-ACC	about	talk-past

(He talked about the flowers in the garden.)

Here the RP verb is ' ulla' . It is followed by ‘pookkaLe’ (NP) and 'patti' (PSP).

Rule 2.If the current token is NP and previous one is an RP verb and if the succeeding token is not PSP then current NP token is marked as clause end.

Example 3:

Aa	valiya	veedu	enikku	venam.
That	big-RP	house	I-DAT	want.

(I want that big house.)

Rule 3. The RP clause can also have RP verb followed by PSP without NP in between.

Example 4:

Avide	poya	shesham	ayaL	ivide	vannu.
There	go-past+RP	after	she	here	come-PAST

(He came here after he went there.)

The grammatical rules will work as follows

If the current token is np,the previous is RP verb and next word is not a PSP then the current np is marked as probable RP clause end.

-1 VM+RP=1

0 NP=1 RP clause end

1 PSP=0

If the current token is a PSP,the previous is a RP verb then current PSP is the probable RP clause end.

-1 VM+RP=1

0 PSP=1 RP clause end

If the current token is a Noun followed by a PSP and the previous is a RP verb then current PSP is the probable RP clause end.

-1 VM+RP=1

0 NP=0

1 PSP=1 RP clause end

2.3.2 Conditional Clauses

There are of two types of conditional clauses:

(1) purely conditional and

(2) hypothetically conditional.

The purely conditional clause will take the morpheme "-a:l" as the suffix of the verb.

Example 5:

nee	nallavannam	padicha:l	tiirchayayum	passakum
you	well	study-COND	surely	pass-FUT

(If you study well,you will surely pass.)

Here the embedded clause is “nee nallavannam padicha:l”

The hypothetical one will take “enkil” as clause conditional particle.

Example 6:

nii	atu	cheyyumenkil	njyan	varaam
you	it	do-FUT-COND	I	come-FUT-MOD

(If you will do it,I will come.)

Beginning of the clause is the first subject NP preceding the VP containing the conditional marker.

The conditional clause ending is found with the rule given below.

If the current verb has a conditional marking suffix, then the current verb is marked for probable conditional clause end.

0 VM+CON=1 CON clause end

Once the innermost clause start is identified the rules are being implemented and then it is repeated until all the different clauses gets their boundaries. We have marked the RP clause start with the value 3 and clause end with -3.For Conditional it is 2 and -2 respectively.

An Example for RP clause handling can be considered.

Example 7:

Ushnakaattu	veeshunna	karnatakayilninn	avan	wayanattil	ethi
Hotwind	blow-PRES-RP	karnataka-LOC-PSP	he	wayanad-LOC	reach-PAST

(From hotwind blowing karnataka, he reached wayanad.)

We do the preprocessing for the part-of-speech and chunking information, and analyze the words with morphanalyser .On the preprocessed text the noun phrase is replaced with np and the head noun morphological information is maintained. The other outputs are altered for better representation in the input to the clause identifier engine.

The altered input is shown below.

np	np	n_nom
vISunna	VM_RP	V_RP
karZNAtakayilZninnu	PSP	I-NP
np	np	n_nom
np	np	n_nom
ethi	VM_VGF	VM_VGF
.	SYM	I-VGF

To the altered input the column representing the rules described above is added. The numbers in the column represent the probable clause start and end marking. Here 3 stands for probable relative participle clause start and -3 to for probable RP clause end. Similarly 6 is for MCL start and -6 is for MCL end.

np	np	n_nom	3
vISunna	VM_RP	V_RP	
karZNAtakayilZninnu	PSP	I-NP	-3
np	np	n_nom	6
np	np	n_nom	
ethi	VM_VGF	VM_VGF	-6
.	SYM	I-VGF	

Same procedure can be followed for Conditional clauses also.

Example 8:

Melle natannu kayariya:l ksheenam ariyilla

(If (you) walk slowly (you) will not feel tired.)

Here 2 stands for probable relative participle clause start and -2 to for probable RP clause end. Similarly 6 is for MCL start and -6 is for MCL end.

np	np	n_nom	2	{CON}	{CON}
np	np	n_nom	o	o	o
kayarYiyAlZ	VM_COND	V_COND	-2	{/CON}	{/CON}
np	np	n_nom	6	{MCL}	{MCL}
np	np	n_nom	-6	{MCL}	{MCL}
.	SYM	I-NP	o	o	o

3 EVALUATION AND RESULTS

We have taken 3638 sentences from tourism corpus and training to testing ratio was about 80% to 20%.We have tagged the sentences for Relative Participle Clauses, Conditional and Main clauses. We trained 2837 sentences and tested the system with 801 sentences.We have used the tags {RP} and {/RP} to mark the RP clause start and end and similarly {CON},{/CON} for Conditional clause start and end and {MCL},{/MCL} for Main clause start and end respectively.The sentences was first preprocessed for POS and chunking information and the words were morphological analyzed. The data is present in column format, with the words forming the first column, pos tags forming the second column, chunking information forming the third column, Boolean entries which obey the linguistic rules forms the fourth column and finally the fifth column is the clause boundary information. The training data of CoNLL had clause information on the fourth column, since we had to add the linguistic feature to the CRF module we used the fourth column for Boolean entries. The Evaluation of the system is given in Table1.

Clauses	Actual	Correct	Tagged	Recall	Precision
RP(Open)	737	686	713	93.08	96.21
RP(Close)	578	380	483	65.74	78.67
CON(Open)	69	56	56	81.16	100
CON(Close)	76	67	68	88.16	98.53
MCL(Open)	287	209	314	72.82	66.56
MCL(Close)	514	478	543	92.99	88.03

Table1: Evaluation of the system.

4 ERROR ANALYSIS

For analysis of erroneous clause boundary marking done by the CRF, the training data was given for testing to the CRF system. From the results obtained it was noted that the RP clause end was not tagged properly when proper nouns with co-ordination marker was encountered.

Example 9:

avide nilkkunna balanum krishnanum aanu ente koottukar

there stand-RP Balan and Krishnan is my friends

(Balan and krishnan standing there are my friends.)

Also cases when 2 RP verbs where coming in succession.

Example 10:

avide kaanunna karangunna silpam.

There see-RP rotate-RP statue

(The rotating statue seen there.)

Conclusion

The system thus developed is the first automatic clause identifier in Malayalam .From the results it was shown that Conditional tags were more accurate. There was more number of RP tags in the starting of many sentences and it was observed that the correctness of RP opening tags was more than closing tags.RP opening tags had a precision of 96.21% and RP close tags had 78.67% precision. Reverse was the case for Conditional and MCL tags. Here we have tried using grammatical rules as one of the feature in Conditional Random Fields.

References

Carreras, X., Marquez L.(2003).Phrase recognition by filtering and ranking with per-ceptrons. *Proceedings of RANLP-2003*, Borovets, Bulgaria. pages 205– 216.

Carreras, X., Marquez L., Punyakanok V., and Roth D.(2002).Learning and inference for clause identification. *Proceedings of the 14th European Conference on Machine Learning,* Finland, pages 35–47

Ejerhed,E.(1988).Finding clauses in unrestricted text by finitary and stochastic methods. *Proceedings of the2nd conference on applied natural language processing*, Austin, Texas, pages 219 – 227.

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

Harris, V.P. (1997).Clause Recognition in the Framework of Alignment, Mitkov, R., Nicolov, N. (Eds.),*Recent Advances in Natural Language Processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pages 417-425.

Kudo,T.(2005). CRF++, an open source toolkit for CRF*, http://crfpp.sourceforge.net.*

Lafferty, J.D., McCallum, A., and Pereira, F.C.N.(2003).Conditional Random Fields: Probabilistic Models For Segmenting and Labeling Sequence Data.*Proceedings of the Eighteenth International Conference on Machine Learning,* pages 282-289.

McCallum,A. and Li,W.(2003).Early results for named entity recognition with conditional random fields, feature induction and web enhanced lexicons. *Proceedings of CoNLL-2003, Edmonton,* Canada, pages 188–191.

Molina, A., Pla, F.(2002).Shallow Parsing Using Specialized HMMs. *Journal of Ma-chine Learning Research 2,* pages 595–613.

Nguyen, V., et. al.(2007). Using Conditional Random Fields for Clause Splitting. *Proceedings of the Pacific Association for Computational Linguistics,* University of Melbourne, Australia .

Orasan,C.(2000). A hybrid method for clause splitting. *Proceedings of ACIDCA 2000 Corpora Processing,* Monastir, Tunisia, pages 129 – 134.

Papageorgiou,H.V.(1997).Clause recognition in the framework of alignment.*Proceedings of Recent Advances in Natural Language Processing, John Benjamins,Publishing Company*, Amsterdam/Philadelphia, pages 417-425.

Parveen, D. , Ansari,A. and Sanyal,R.(2011).Clause Boundary Identification using Clause Markers and Classifier in Urdu Language.*12th International Conference on Intelligent Text Processing and Computational Linguistics CICLing* .

Puscasu,G.(2004).A Multilingual Method for Clause Splitting. *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, Birmingham, UK.

Sha,F. and Pereira,F.(2003).Shallow parsing with conditional random fields.*Proceedings of HLT- NAACL03*,pages 213–220 .

Sobha,L. and Patnaik,B.N.(2002).Vasisth: An anaphora resolution system for Malayalam and Hindi. *Symposium on Translation Support Systems.*

Tjong,E. ,Sang ,K. and Dejean ,H.(2001).Introduction to the CoNLL-2001 shared task: clause identification, *Proceedings of the 2001 workshop on Computational Natural Language Learning* ,Toulouse,France.

Vijay Sundar Ram R., Bakiyavathi T. and Sobha L. (2009). Tamil Clause Identifier. *PIMT Journal of Research, Patiala, Vol.2. No.1,* pages 42-46.

Vijay Sundar Ram R. and Sobha Lalitha Devi.(2008). Clause Boundary Identification Using Conditional Random Fields. *In Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 4919/2008*,pages 140-150 .

Vilson J. Leffa(1998). Clause processing in complex sentences. *Proceedings of the First International Conference on Language Resource & Evaluation,*pages 937 – 943 .