ISSN No. 0976-5697

Volume 7, No. 4, July-August 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Clause Boundary Identification for Non-Restrictive Type Complex Sentences in Telugu Language

V. Suresh
Department of Computer Science and Engineering,
Vignan's Institute of Information Technology, Duvvada,
Visakhapatnam, A.P., India

M.S. Prasad Babu
Department of Computer Science and Systems Engineering,
Andhra University College of Engineering,
Andhra University, Visakhapatnam, A.P., India

Abstract: Sentence identification for non-restricted type Telugu language is one of the basic applications of the Natural Language Processing. A sentence composed of single independent clause is called a simple sentence. If a sentence contains a dependent clause along with independent clause then it is a complex sentence. Once the sentence is identified as complex sentence, the next step is to identify its pattern. After identification of patterns, various clauses present in the sentence are extracted and grammar checking is performed on them. For grammar checking of complex sentences, it is necessary to identify the structure of Clause Boundary for Non-Restrictive Type of Telugu Language Complex Sentences. In this paper we have explored the different types of sentences present in Telugu language. The structure of complex sentences can be identified on the basis of number of clauses and types of clauses present in them. We have also proposed an algorithm for identification of simple, compound and complex sentences. This study will be helpful in identifying and separating the complex sentences from Telugu language. We have also proposed an algorithm for identification of simple, compound and complex sentences. Also this study will be helpful in developing other Natural Language Processing (NLP) applications like converting a complex sentence in simple sentences, grammar checking of complex sentences.

Keywords: NLP, Complex Sentences, Independent clause, Dependent clause.

INTRODUCTION

A clause is a grammatical unit that includes, at minimum, a predicate and an explicit or implied subject, and expresses a proposition. Larger sentences (compound and complex) are composed of more than one dependent clause. The performance of grammar- checker application decreases as the complexity or size of sentences increases.. (Loos, Anderson, Day, Jordan & Wingate, 1999). The clauses present in a sentence can be of same type (independent clauses) or of different types (dependent and independent). A complex sentence cannot be processed until all its clauses are not properly identified (Leffa, 1998). Clause boundary identification is one of the shallow semantic parsing tasks consisting of marking the starting and end position of a clause in the sentence. Clause boundary identification is a special kind of shallow parsing, like text chunking. It is more difficult than phrase chunking, since clauses can have embedded clauses. Due to the ambiguous nature of natural languages by using Clause boundary identification of natural language sentences poses considerable difficulties.

Telugu language belongs to Indo-Aryan family of languages (Dravidian languages). Other members that belong to this family are Kannada, Tamil, Malayalam, Hindi, Bengali, Gujarati, and Marathi etc. Telugu is spoken in India, Canada, USA, UK, and other countries with Telugu immigrants. Telugu language is the 8th most widely spoken language in the world, 4th most spoken language in Canada (The Times of India, 14th February, 2008) and the 9th in India with more than 45 million speakers. It is the official language of Telugu states (Andhra Pradesh and Telangana).

The first treatise on **Telugu grammar**, the "Andhra Shabda Chintamani" was written in Sanskrit by Nannaya who was considered as the first poet and translator of Telugu in the 11th century A.D. There was no grammatical work in Telugu prior to Nannayya's "Andhra shabda chintamani". This grammar followed the patterns which existed in grammatical treatises like Astadhvavi and Valmiki vyakaranam but unlike Paninni. Nannayya divided his work into five chapters, covering Samjna, Sandhi, Ajanta, Halanta and Kriya. After Nannayya, Atharvana and Ahobala composed Sutras, Vartika and, Bhashyam. Like Nannayya, they had previously written their works in Sanskrit.

In the 19th century, Chinnaya Suri wrote a simplified work on Telugu grammar called Balavyakaranam borrowing concepts and ideas from Nannayya's Andhra Shabda Chintamani, and wrote his literary work in Telugu. Every Telugu grammatical rule is derived from Paninian, Katvayana, and Patanjali concepts. However high percentage of Paninian aspects and techniques borrowed in Telugu.

According to Nannayya language without 'Niyama'' or the language which doesn't adhere to Vyākaranam is called Gramya or Apabramsa and hence it is unfit for literary usage. All the literary texts in Telugu follows Vyākaranam. Compared to languages like English, Telugu is a morphologically rich language and has relatively free word order. It follows a **Subject-Object-Verb** (S-O-V) pattern. It is:

Sentence	బాలుడు పాఠశాలకు బయల్దేరాడు		
Words	బాలుడు	పాఠశాలకు	బయల్దేరాడు
Transliteration	Baludu	pataselaku	bayalderadu
Gloss	Boy	towards the school	moved
Parts	subject	object	verb
Translation	Boy moved towards the school		

This sentence can also be interpreted as 'Boy moved towards the school' depending on the context. But it does not affect the SOV order.

1.1 CLAUSES:

All the phrases (postpositional, nominal, adjectival, verb) combines to constitute the clauses. If a sentence is highest unit, then clause is the second highest unit of the sentence. These are composed of phrases.. A clause may contain any number of phrases. Verb phrase is the essential component of every clause. Even a single clause constituted by a verb phrase can construct sentence. There is no need of any other element in the sentence.

Clauses can be classified on the basis of these verb phrase; The verb phrase is the essential element of every clause. There are two types of clauses in Telugu language one is independent and other is the dependent clause. The clause having the finite verb phrase is called independent clause and the other having non-finite verb phrase is called dependent clause. In the following section, an overview of these two types of clauses are given.

1.1.1 INDEPENDENT CLAUSES:

Independent clause is essential part of all types of sentences. The structure of independent clause is same as that of simple sentence. A clause is called independent clause if it can exist independently as a complete sentence. Verb phrase is the essential part of the independent clause. The independent clause contains exactly one verb phrase (Puar, 1990) along with other elements of the clause. Other than verb phrase, an independent clause may contain one or more noun phrase, adjective phrase, adverb phrase etc. as other elements (Bray, 2008). The verb phrase present in the independent clause is finite verb phrase. In compound sentences, these independent clauses are joined by coordinate conjunctions. In complex sentences, an independent clause and dependent clause are joined using subordinate conjunctions. Independent clauses are used to give more identity in grammar checking system.

1.1.2 DEPENDENT CLAUSES:

Subordinate verb phrase as one of the essential element of type of independent clause. These clauses convey an incomplete thought and hence, cannot constitute a sentence without the help of other clauses. To constitute a sentence it requires at least one independent clause. These clauses participate in the construction of complex sentences. Dependent clause also contains verb phrase as one of its essential element .like independent clause, This verb phrase can be finite or non-finite.

II. LITERATURE SURVEY

Poornima C. Dhanalakshmi V.Anand Kumar M.Soman KP (2011) have developed Rule Based sentence Simplification for English to Tamil Machine Translation System. They presented the simplification of complex sentence in English language. They explained how to convert a complex sentence to simple sentence and also translate them to Tamil language without changing them meaning of the sentence. Machine Translation was the process by which the computer software is used to translate a text from one natural language to another but handling complex sentences by any machine translation system was considered to be difficult. In this paper rule based technique was used to simplify the complex sentences. This technique was based on connectives like relative pronouns, coordinating and subordinating conjunctions to obtain simple sentences for machine translation. Sentence simplification, Sentence segmentation, POS tag and Machine translation were used in this paper. In Machine Translation system 100% accuracy was not possible. **Kamaljeet Kaur Batra and GS Lehal** have developed the Rule Based Machine Translation of Noun Phrases from Punjabi to English. They presented the automatic translation of noun phrases from Punjabi to English using transfer approach. Preprocessing tagging, ambiguity resolution, translation and synthesis of words in target language were the various steps that were involved in this paper. They used the Morphological Analyzer tool for translation and used rule base technique. Accuracy was calculated for each step. The overall accuracy of the system was calculated to be about 85% for a particular type of noun phrases.

NaushadUzZaman, Jeffrey P. Bigham and James F. Allen (2011) proposed a rule based system for the simplification of the sentences. This simplification was required to improve the machine translation system. The machine translation system from English to Tamil was developed by the authors. This system lacks in accuracy because of problem in translating compound and complex sentences from English to Tamil language. To overcome this difficulty they proposed a system that will first identify the compound and complex sentences and then simply convert them to simple sentences. Handmade rules were used to develop this system.

Daraksha Parveen, Ratna Sanyal and Afreen Ansari have developed Clause Boundary Identification using Classifier and Clause Markers in Urdu Language. They presented the identification of clause boundary for the Urdu language. They used Conditional Random Field as the classification method and the clause markers. The clause markers play the role to detect the type of subordinate 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.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.

.

III. NON-RESTRICTIVE TYPE TELUGU LANGUAGE COMPLEX SENTENCES

Nonrestrictive clauses provide some information about the preceding subject and conjunctions starting with ` ో'(llo) character are used to provide relevant information of the clause. Comma is used to separate different clauses. Consider the following examples:

S.No	Example
1	కొన్ని వస్తువుల్లో అవి ఎలా అమర్చారంటే , చూడడానికి అక్కడే వుండాలనిపిస్తుంది (konnivasthuvulloavielaamarcharante,chudadanikiakkade undalanipisthundi) A few objects,which have been placed , look lively
2	లుమగలిద్దరూ, ఆఫీసుకి వెళ్ళి, ఇప్పుడు ఇంటికి వచ్చారు. (aalumagaliddaru, office ki, velli, ippuduintikivaccharu.) Both wife and husband, went to office, return to home.

Like restrictive clause, this clause is also embedded between the subject and the predicate of the independent clause. Therefore, like restrictive clause, nonrestrictive clause also splits the sentence in three parts. One part containing start and end of the subject of independent clause, second part containing the start and end of the nonrestrictive clause and the third part containing the start and end of the predicate of independent clause.

Consider the sentence 1 from above table:

కొన్ని వస్తువుల్లో అవి ఎలా అమర్చారంటే, చూడడానికి అక్కడే వుండాలనిపిస్తుంది

(konnivasthuvulloavielaamarcharante,chudadanikiakkadeundalanipisthundi)

A few objects,which have been placed , look lively

As shown in the above complex sentence, the nonrestrictive clause అవి ఎలా అమర్చారంటే☎avi ela amarcharante) is embedded between the subject కొన్ని వస్తువుల్లో ☎konni vasthuvullo) and the predicate చూడడానికి అక్కడే వుండాలనిపిస్తుంది (chudadaniki akkade undalanipisthundi) of the independent clause. In this way, nonrestrictive clause splits the sentence into three parts. The first part contains the subject of independent clause, the second part contains the nonrestrictive clause and the third part contains the predicate of independent clause. The first part of the sentence starts with the first word of the sentence i.e. కొన్ని ☎konni) and ends with the word having comma and just previous to తే ☎the) conjunction i.e. with the word ల్లో(llo) . Similarly, the second part of the sentence starts with the తే ☎the) conjunction that appears just after the first comma and ends with the word just before the second comma i.e. రంటే ☎rante) , and the third part starts with a word just after the second comma i.e. చూడడానికి ☎chudadaniki) and ends with the last word of the sentence. These three separate parts of the sentence can be represented as:

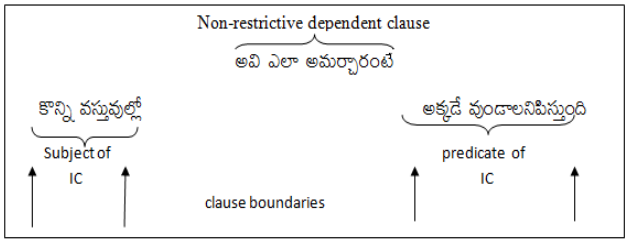

Figure 3.1: Marking clause boundaries in complex sentence having non-restrictive type clause

As shown above, the start and end point of the nonrestrictive clause can be identified with తే ☎the) conjunction and comma (,) respectively. తే ☎the) conjunction separates the subject of independent clause from nonrestrictive clause and common restrictive clause from the predicate of dependent clause.

Algorithm:

Algorithm used: Identification of Clause boundary of non-predicate bound nonrestrictive type.

Input: Annotated Telugu sentence

Database used: List of తే ☎the) conjunction

Output: Telugu sentence with marked clause boundaries.

1. Tokenize the input sentence.
2. Mark the first word of the sentence as beginning of subject of independent clause.
3. Repeat step 4 for all the tokens of the sentence.
4. If the current word is comma and the next word is తే × ☎the) conjunction then go to step 5.
5. Mark the previous word as end of subject of independent clause and go to step 6.
6. Mark the next word i.e. తే ☎the) conjunction as beginning of dependent clause.
7. If current word is auxiliary verb with comma and it is not the last word of the sentence then mark this word as end of dependent clause and next word as beginning of predicate of independent clause.
8. Mark the last word as end of predicate of independent clause.

Flow chart representing above mentioned algorithm is shown in figure 5.2. An example to illustrate the working of flowchart/algorithm is also provided with this flow chart.

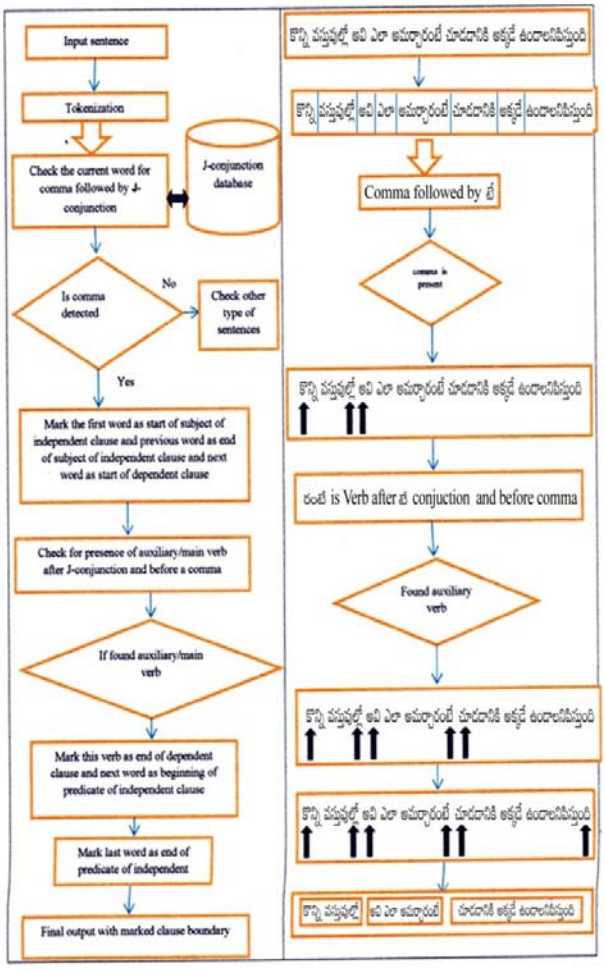

Fig 3.2: Flow chart and working example to mark clause boundaries in nonrestrictive type complex sentences

RESULTS AND DISCUSSION

4.1. COMPLEX SENTENCES OF GRAMMATICAL MISTAKES:

In this system, the grammatical errors in complex sentences have been detected at clause level (independent clause) and sentence level (various agreements between dependent and independent clauses). Besides these, some other clause level and sentence level grammatical errors in complex sentences detected and corrected by this system have been listed in table 4.1.

S.No	Type of error	Example
1.	Use of first form of verb before 4 particle in dependent clause.	Incorrect: నేలు పాలు త్రాగుతూ స్కూలుకి వెళ్ళిపోతాను (nenu palu traguthu skoolki vellipothanu) Correct: నేలు పాలు త్రాగి స్కూలుకి వెళ్ళిపోతాను nenu palu tragi skoolki vellipoothanu)
2.	Noun phrase and verb phrase agreement between dependent and independent clauses.	అబ్బాయి స్కూలుకి వెళుతూ వుంటే ఖచ్చితంగా పాస్ అవు (abbayi skoolki veluthoo unte khacchithanga Pass avu) అబ్బాయి స్కూలుకి వెళుతూ వుంటే ఖచ్చితంగా పాస్ అవుతాడు (abbayi skoolki veluthoo unte khacchithanga Pass avuthadu)
3.	Noun phrase and verb phrase agreement within independent clause.	అతడు బాగా కష్టపడినా కూడా పరీక్షల్లో నెగ్గక పోయినారు (athadu baga kashtapadina kudapareeshalo pass avakapoyinaru) అతడు బాగా కష్టపడినా కూడా పరీక్షల్లో నెగ్గక పోయినాడు (athadu baga kashtapadina Kudaa pareeshalo pass neggaka poyadu)
4.	Common subject and verb Agreement	Incorrect నా తమ్ముడు మంచివాడు మరియు రోజూ ఇంటి వస్తాడు. (naa thammudu manchivaadu mariyu roju inti vasthadu) Correct నా తమ్ముడు మంచివాడు మరియు రోజూ ఇంటికి వస్తాడు. (naa thammudu manchivaadu mariyu roju intiki vasthadu)

Table: 4.1: Various types of errors detected in complex sentences

We tested our module on Telugu randomly picked from the internet. We take two samples from different sites.
One sample is given name set A and the second Sample given name set B.

Test Set	Size (No. of Sentences)	Accuracy	
		Predicate Bound	Non-Predicate Bound
A	2400	85%	81%
B	3100	82%	80%

CONCLUSION

This paper concerns the grammar checking of complex sentences in various agreement errors within independent clauses in case of complex sentences and between dependent and independent clauses in case of complex sentences. For grammar checking of complex sentences, it is necessary to identify the structure of Clause Boundary for Non-Restrictive Type of Telugu Language Complex Sentences. In this paper we have explored the different types of sentences present in Telugu language. The structure of complex sentences can be identified on the basis of number of clauses and types of clauses present in them. We have also proposed an algorithm for identification of simple, compound and complex sentences. This study will be helpful in identifying and separating the complex sentences from Telugu language. We have also proposed an algorithm for identification of simple, compound and complex sentences. Also this study will be helpful in developing other Natural Language Processing (NLP) applications like converting a complex sentence in simple sentences, grammar checking of complex sentences.

V. REFERENCES

1. Sobha, L. D., & Lakshmi, S. Malayalam. 2013. Clause Boundary Identifier: Annotation and Evaluation. *WSSANLP-2013, p.* 83.
2. Kaur, N., Garg, K., Sharma, Sanjeev. Kumar. 2013. Identification and Separation of Complex Sentences from Punjabi Language. *International Journal of Computer Applications,* 69(13), pp. 21-24.
3. Sharma, Sanjeev Kumar **'**Assigning the Correct Word Class to Punjabi Unknown Words using CRF' *International Journal of Computer Applications (0975 – 8887) Volume 142 – No.2, May 2016*
4. Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language.* Association for Computational Linguistics. pp. 112-116
5. Brill, E. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In *Proceedings of the workshop on Human Language Technology.* Association for Computational Linguistics. pp. 237242
6. Kasbon, R., Amran, N., Mazlan, E., & Mahamad, S. 2011. Malay language sentence checker. *World Appl. Sci. J. (Special Issue on Computer Applications and Knowledge Management), 12, pp.* 19-25.
7. Kaur, N., Garg, K., & Sharma, S. K. 2013. Identification and Separation of Complex Sentences from Punjabi Language. *International Journal of Computer Applications,* 69(13), pp. 21-24.
8. Kubon V., & Platek, M. 1994. A grammar based approach to a grammar checking of free word order languages. In *Proceedings of the 15th conference on Computational linguistics-Volume 2.* Association for Computational Linguistics. pp. 906-910
9. Leffa, V. J. 1998. Clause processing in complex sentences. In Proceedings *of the First International Conference on Language Resources and Evaluation* Vol. 1, pp. 937-943.

10. Narula, R., & Sharma, S. K. 2014. Identification and Separation of Simple, Compound and Complex Sentences in Punjabi Language. International Journal of Computer Applications & Information Technology. Vol. 6, Issue II Aug-September 2014.
11. Orasan, C. 2000. A hybrid method for clause splitting in unrestricted English texts. *Proceedings of ACIDCA' 2000*
12. Parveen, D., Sanyal, R., & Ansari, A. 2011. Clause Boundary Identification using Classifier and Clause Markers in Urdu Language. Polibits *Research Journal on Computer Science,* 43, pp. 61-65.
13. http://en.wikipedia.org/wiki/Telugu_languag.
14. http://en.wikipedia.org/wiki/Telugu_grammar.
15. http://simple.wikipedia.org/wiki/Telugu_language.