JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Marking Clause Boundary in Compound Sentences of Punjabi Language

S. K. Sharma

Dept. of Computer Science and Applications, DAV University, Jalandhar (Punjab), India

**Corresponding Author: sanju3916@rediffmail.com*

Available online at: www.ijcseonline.org

Received: 22/Aug/2017, Revised: 08/Sep/2017, Accepted: 19/Sep/2017, Published: 30/Sep/2017

***Abstract*---**Clause boundary identification for compound sentences in Punjabi language is one of the basic necessity for processing of compound sentences. For grammar checking of compound sentences, it is necessary to identify the structure of various independent clauses present in compound sentence. Once the sentence is identified as compound sentence, the next step is to identify its pattern. After identification of patterns, various clauses present in the sentence are extracted as it is the basic step for performing grammar checking. In this paper, author has explored a technique to identify the clause boundaries present in compound sentence. This study will be helpful in identifying and separating the compound sentences from Punjabi language corpus. Also this study will be helpful in developing other Natural Language Processing (NLP) applications like simplification compound sentence in simple sentences, Improving Machine translation system and grammar checking of compound sentences.

***Keywords*--**NLP, Compound Sentences, Independent clause, grammar checking.

I. INTRODUCTION

A clause is a largest unit of the sentence that contains a predicate and an explicit or implied subject. A sentence may have any number of clauses. Clause boundary identification means to split the sentence into clauses by identifying the starting and the ending position of the clause. The task of clause boundary identification is not only detecting a non-recursive phrase of the sentence, rather it is a three step process: identifying start of clause, identifying end of clause and finding complete clause (Sang and Dejean, 2001). Consider the following example:

Punjabi: ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ ਤੇ ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ।

Transliteration: (mīṃh pai rihā sī lōk bhijj rahē san)
Translation: It was raining and people were on

In the above sentence, there are two clauses; one is ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ (mīṃh pai rihā sī) and second is ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ (lōk bhijj rahē san). Both these clauses are joined by conjunction (ਤੇ). In clause identification, problem to be identified is the start and end position of both the clauses. As shown in fig 1, s1 and e1 represent start and end point of the first clause (clause 1) and s2 and e2 represent start and end point of the second clause.

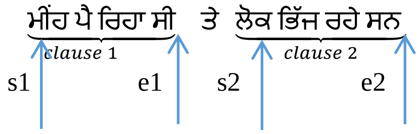

Fig 1: Marking starting and end points of clauses in a sentence

II. 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 Punjabi 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.

III. 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.

IV. COMMON EXISTING APPROACHES

Work on clause identification has been done on a few Indian languages and some foreign languages. The common approach followed for most of the clause identification systems can be categorized into two types – rule based and statistics based. Hybrid of these two approaches has also been attempted. Conditional Random Field as the classification method and the clause markers was proposed for Urdu language (D. Parveen et al., 2009). Aniruddha Ghosh et al. (2010) proposed a hybrid approach (combination of rule based and statistics based approach) for development of clause identification and separation system in Bengali language. Conditional Random Fields (CRFs) framework was used for clause splitting problem in English language (Vinh Van Nguyen et al., 2009). A rule based technique was followed for identifying clause boundaries in text (Papageorgiou, 1997). Rule based approach was used for English (Leffa, 1998). A memory-based learner with post-processing rules was used for predicting clause boundaries in Susanne corpus (Orasan, 2000). Georgiana Puscasu (2004) proposed a multilingual method for detecting clause boundaries in unrestricted texts. Erif K.Tjong et al. (2001) proposed a memory based learner to CONLL-2001 shared task. Erik F. Tjong et al. (2001) proposed a machine learning system for identification of clauses. Eraldo Frenandes et al. (2009) proposed Entropy guided transformation learning (ETL) method for clause identification. Zhemin Zhu et al. (2010) proposed a statistical based sentence simplification model for converting the complex sentences into simple sentences. Ani Thomas et al. (2011) used dependency relationship for identification and separation of clauses in a sentence. Naushad UzZaman et al. (2011) proposed a rule based system for the simplification of the sentences. Xavier Carreras et al. (2001) used Ada Boost Learning Algorithm to solve the simple decision for clause splitting problem.

V. STRUCTURE OF INDEPENDENT CLAUSES

A clause is called independent clause if it can exist independently as a complete sentence. The structure of independent clause is same as that of simple 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 (Brar, 2008). The verb phrase present in the independent clause is finite verb phrase. Independent clause is essential part of all types of sentences. 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.

A. Independent clause as a simple sentence:

In simple sentence, there is only one clause and this clause is independent clause. For example:

$$\underbrace{\text{ਉਹ}}_{noun\ phrase} \quad \underbrace{\text{ਦਿੱਲੀ}}_{noun\ phrase} \quad \underbrace{\text{ਗਿਆ}}_{verb\ phrase}$$

(uh dillī giā)

Above sentence contain one verb phrase ਗਿਆ (giā) along with other two noun phrases.

B. Independent clauses in compound sentences

A compound sentence can have more than one independent clauses. Consider the following example:

ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ ਤੇ ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ

(mīṃh pai rihā sī te lōk bhijj rahē san).

$$\underbrace{\text{ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ}}_{Independent\ clause1} \quad \underbrace{\text{ਤੇ}}_{conjunction} \quad \underbrace{\text{ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ}}_{independent\ clause\ 2}$$

Above compound sentence contains two independent clauses. First is ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ (mīṃh pai rihā sī) and second is ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ (lōk bhijj rahē san).

*C. **Independent clauses in complex sentences***

A complex sentence can have an independent clause along with dependent clause. For example:

ਰੋਟੀ ਖਾ ਕੇ ਬੱਚੇ ਖੇਡਣ ਚਲੇ ਗਏ

rōṭī khā kē baccē khēḍaṇ calē gaē

$\underbrace{\text{ਰੋਟੀ ਖਾ ਕੇ}}_{\text{dependent clause}}\ \underbrace{\text{ਬੱਚੇ ਖੇਡਣ ਚਲੇ ਗਏ}}_{\text{independent clause}}$

In above complex sentence, there is one independent clause i.e. ਬੱਚੇ ਖੇਡਣ ਚਲੇ ਗਏ (baccē khēḍaṇ calē gaē)

VI. IDENTIFICATION OF INDEPENDENT CLAUSES

An independent clause can itself be a sentence (simple sentence) or it can be part of larger sentence (compound or complex sentence). In simple sentence, there is only one clause and therefore, independent clause can be identified by marking the starting and end position of the sentence. On the other hand, compound sentences are composed of more than one independent clauses joined by coordinate conjunctions. These coordinate conjunctions can be used to mark the start and end position of independent clauses. Consider the following example:

Punjabi: ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ ਅਤੇ ਉਹ ਮੇਰੇ ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ

Transliteration: (ih mērā sagā bharā hai atē uh mērē cācē dā muṇḍā hai)

In the above compound sentence, there are two independent clauses; one is "ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ" (ih mērā sagā bharā hai) and second is "ਉਹ ਮੇਰੇ ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ" (uh mērē cācē dā muṇḍā hai). Both these clauses are joined by coordinate conjunction ਅਤੇ (atē). This coordinate conjunction ਅਤੇ (atē) can be used to mark the end position of the first independent clause and start position of the second independent clause. Above sentence, after marking the independent clauses, can be viewed as:

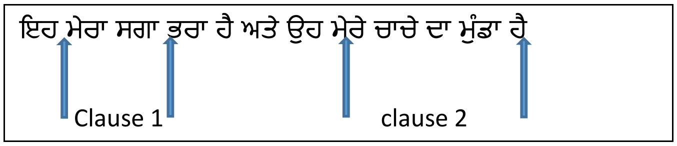

Figure 2: Compound sentence with marked clause boundaries

Algorithm used: Clause boundary identification of independent clauses in compound sentence.

Input: Annotated Punjabi compound sentence.

Database used: List of coordinate conjunctions.

Output: Punjabi sentence marked with clause boundaries.

- Tokenized the input sentence.
- Mark the first word of the sentence as beginning of independent clause.
- Repeat step 4 for all the tokens of the sentence.
- If the current word is coordinate conjunction and is not a part of any phrase, then go to step 5.
- Mark the position of previous word to this conjunction as end of clause and go to step 6.
- Mark the next word to this conjunction as beginning of independent clause.
- Mark the last word as end of independent clause.

Flow chart representing above mentioned algorithm is shown in figure 3. An example to illustrate the working of flowchart/algorithm is also provided with this flow chart.

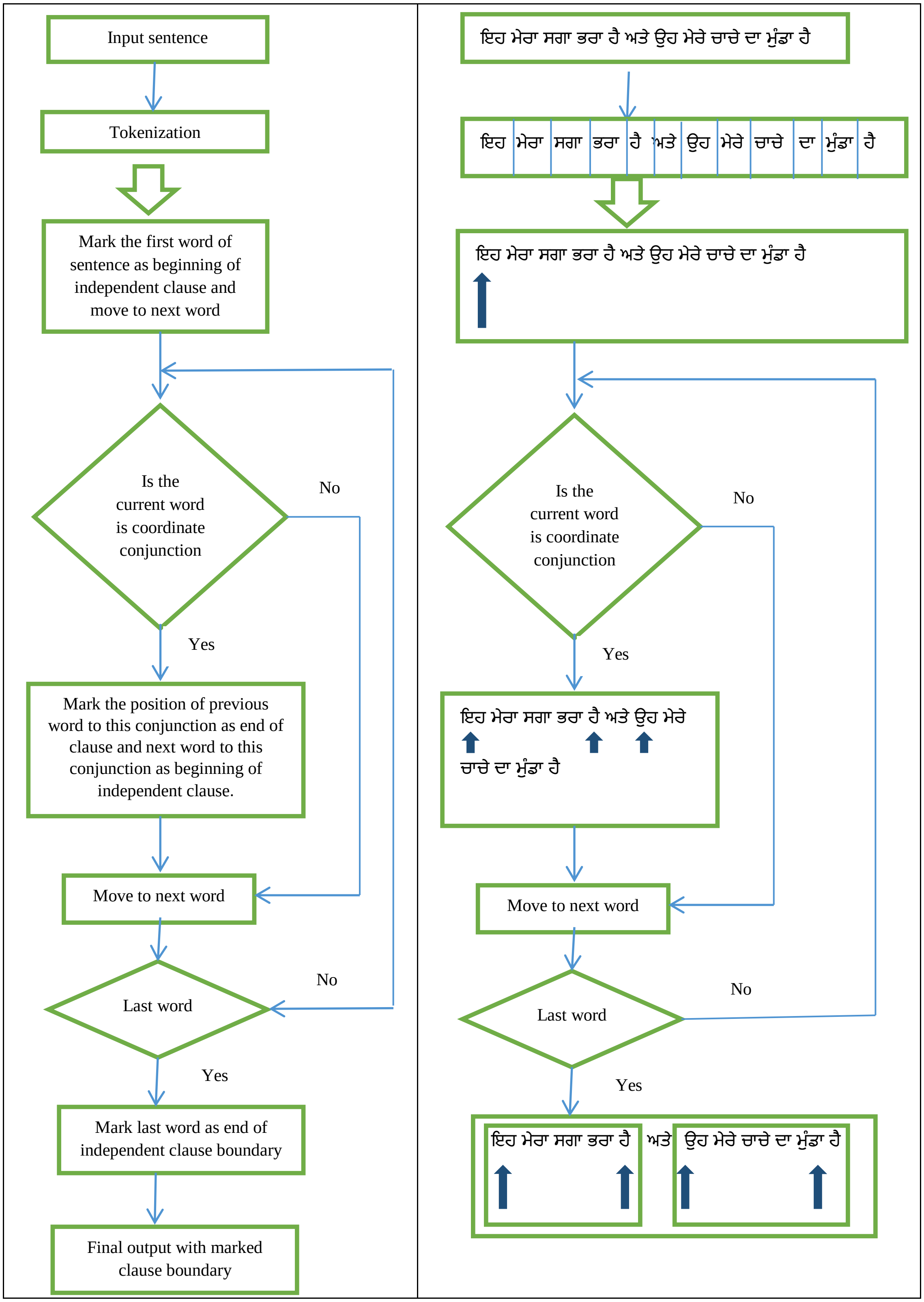
Input sentence
Tokenization
Mark the first word of sentence as beginning of independent clause and move to next word
Is the current word is coordinate conjunction
No
Yes
Mark the position of previous word to this conjunction as end of clause and next word to this conjunction as beginning of independent clause.
Move to next word
Last word
No
Yes
Mark last word as end of independent clause boundary
Final output with marked clause boundary
ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ ਅਤੇ ਉਹ ਮੇਰੇ ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ
ਇਹ | ਮੇਰਾ | ਸਗਾ | ਭਰਾ | ਹੈ | ਅਤੇ | ਉਹ | ਮੇਰੇ | ਚਾਚੇ | ਦਾ | ਮੁੰਡਾ | ਹੈ
ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ ਅਤੇ ਉਹ ਮੇਰੇ ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ
Is the current word is coordinate conjunction
No
Yes
ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ ਅਤੇ ਉਹ ਮੇਰੇ
ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ
Move to next word
Last word
No
Yes
ਇਹ ਮੇਰਾ ਸਗਾ ਭਰਾ ਹੈ
ਅਤੇ
ਉਹ ਮੇਰੇ ਚਾਚੇ ਦਾ ਮੁੰਡਾ ਹੈ

Figure3: Flow chart with working example to mark clause boundary in compound sentences

VII. EXPERIMENTAL RESULTS

Author tested his system on a Punjabi corpus containing 239 simple sentences, 200 compound sentences and 220 complex sentences. In table 1, experimental results in the form of precision, recall and F score are tabulated. Results show that the developed system performs well for all types of sentences.

Table 1: Experimental Evaluation of Identification of independent clauses

Sentence type	**Total no. of independent clauses present in the sentences (A)**	**No. of correctly identified independent clauses (B)**	**No. of incorrectly identified independent clauses (C)**	**Precision** $\frac{B+C}{A}$ **X 100**	**Recall** $\frac{B}{A}$**X 100**	**F score** $\frac{Precision\ X\ Recall}{precision+recall}$**X2**
Simple	239	237	2	100	99.16	99.57823
Compound	200	190	10	100	95.00	97.4359
Complex	220	210	10	100	95.45	97.67204

VIII. COMPARISON WITH THE EXISTING SYSTEM

As shown in table 1, author's proposed system identifies the clause boundaries of independent class with a recall of 99.16 for simple sentences, 95% for compound sentences and 95.45 for complex sentences. Zhou et al. (2010) reported a precision of 73.36% and a recall of 80.02% for Chinese language, Ghosh et al. (2010) reported an accuracy of 73.12% using rule based technique and 78.07 using CRF for Bengali language, Pu.sca.su (2004) reported an accuracy of 95% for Romanian language and 92% for English language, Jorgensen (2007) reported an accuracy of 94.37% for spoken language of Norwegian, Nguyen et al. (2007) shows a precision of 90.01% for English.

IX. CONCLUSION AND FUTURE SCOPE

This paper concerns the marking of independent clauses in compound sentences. It can be concluded from the results shown in table 1 that the purposed system performs well in identification of independent clauses in all the three types of sentences i.e. simple, compound and complex sentences. Further this system can be improved by adding the feature of detection of dependent clauses.

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]. 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

[8]. 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.

[9]. 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.

[10]. Orasan, C. 2000. *A hybrid method for clause splitting in unrestricted English texts*. Proceedings of ACIDCA' 2000

[11]. 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.