Identification and Separation of Complex Sentences from Punjabi Language

Navneet Kaur
Department of CSE/IT
Lovely Professional University
Phagwara, Punjab, India

Kamaldeep Garg
Department of CSE/IT
Lovely Professional University
Phagwara, Punjab, India

Sanjeev Kumar Sharma
Head Of Department (CSE)
BIS College Of Engg. And Tech.
Gagra Moga

ABSTRACT
Complex sentences constitute major parts of the Punjabi language. All the large sentences are either of compound or of complex type. Detail analysis of complex sentences is helpful in processing the Punjabi language through computer. This study will be helpful in identifying and separating the complex sentences from Punjabi corpus. Also this study will be helpful in developing other NLP applications like converting a complex sentence in simple sentences, grammar checking of complex sentences, summarization and machine translation etc.

1. INTRODUCTION
A **complex sentence** is composed of dependent and independent clauses. It must contain at least one independent clause and one or more than one dependent clauses. The dependent and independent clauses are joined by using subordinate conjunctions. Conjunctions are words such as for, and, nor, but, or, yet, so. The structure of compound sentences is symmetrical. This structure is composed of two or more than two independent clauses. These independent clauses are composed joined by co-ordinate conjunctions.

ਜੇ ਤੂੰ ਸਾਡੇ ਨਾਲ਼ ਜਾਣਾ ਹੈ **ਤਾਂ** ਆ ਜਾ ।

In above sentence JE and TAN are sub-ordinate conjunctions. They occurs in pair i.e. one part lies at the starting position and the other part lies at the beginning of an independent clause. In above sentence **ਜੇ** ਤੂੰ ਸਾਡੇ ਨਾਲ਼ ਜਾਣਾ ਹੈ is the dependent clause starting with subordinate conjunction **ਜੇ** and the other part ਆ ਜਾ is the independent clause joined with dependent clause with the second part of sub-ordinate conjunction (**ਜੇ** -**ਤਾਂ**)i.e. **ਤਾਂ .[11]**

2. OVERVIEW OF PUNJABI LANGUAGE
Punjabi language is a member of the Indo-Aryan family of languages, also known as Indic languages. Other members of this family are Hindi, Bengali, Gujarati, and Marathi etc. Indo-Aryan languages form a subgroup of the Indo-Iranian group of languages, which in turn belongs to Indo-European family of languages. Punjabi is spoken in India, Pakistan, USA, Canada, England, and other countries with Punjabi immigrants. It is the official language of the state of Punjab in India. Punjabi is written in „Gurmukhi‟ script in eastern Punjab (India), and in „Shahmukhi‟ script in western Punjab (Pakistan).[11]

3. INTRODUCTION TO CLAUSES
Clauses are second largest unit in a sentence. These are made up by combination of different phrases. There are basically two types of clauses Dependent clause and Independent clause. The basic difference between these two types of clauses is that an independent clause can constitute an independent sentence where as a dependent clause cannot create a sentence, they need an independent clause for it.[11]

3.1 Independent Clause
An independent clause can constitute a simple sentence on its own. Every sentence contains independent clause as a basic element. The independent clause contains a finite verb phrase as an essential element.[11]

3.2 Dependent Clause
Dependent clause cannot constitute a sentence on its own. It always occurs with an independent clause in a complex sentence. It typically consists of subordinate verb phrases or start with a subordinate conjunction or words of relative pronoun word class. Dependent clause typically conveys an incomplete thought and for its completion an independent clause is required. In some cases it can have the structure similar to an independent clause.
Dependent clauses in complex sentences can be realized by the following means:
The complex sentence containing three or more than three clause can have more than one independent clause.
The dependent clauses of complex sentences contain noun phrase, adverb phrase and adjective phrase.[11]

4. LITERATURE REVIEW
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 the 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 conjuctions 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.[1]

Kamaljeet Kaur Batra and GS Lehal have developed the Rule Based Machine Translation of Noun Phrases fom 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.[6]

Naushad UzZaman , 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. Hand made rules were used to develop this system.[7]

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 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.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.[9]

5. STRUCTURE OF PUNJABI SENTENCE

Punjabi sentence follow SOV (Subject-Object-Verb) order. In Punjabi sentences, the subject occurs first followed by the object and then the verb. Punjabi sentences can be categorized in to three types. These are simple sentence, compound sentence and complex sentences. A sentence is further composed of clauses which can be further classified as in-dependent clause and dependent clause. An independent clause can constitute a simple sentence on its own. Every sentence contains independent clause as a basic element. The independent clause contains a finite verb phrase as an essential element.[11]

6. PATTERENS OF COMPLEX SENTENCES

On the basis of arrangement of independent clauses and dependent clauses using subordinate conjunctions following patterns have been designed:-

Pattern 1:-

/ /Independent Clause/ / +/ / Sub-ordinate Conjunction / /+ / /Dependent Clause/ /

In this type of complex sentences one independent clauses and one dependent clause is joined by using sub-ordinate conjunction.

/ /ਵਾਕ ਰਚਨਾ ਤਾਂ ਹਰ ਕੋਈ ਕਰ ਸਕਦਾ ਹੈ / /ਜੇ/ / ਉਸਨੂੰ ਅਕਲੀ ਹੁਨਰ ਹੋਵੇ ।/ /

Pattern 2:-

/ /Sub-ordinate Conjunction / /+ / /Dependent Clause/ / +/ / Independent Clause/ /

This type of complex sentence starts with sub-ordinate conjunction followed by a dependent clause and then further followed by an in-dependent clause.

// ਜਿਉਂ ਜਿਉਂ // ਨਵੀਂ ਤਹਜ਼ੀਬ ਵੱਧ ਰਹੀ ਹੈ // ਖੁਸ਼ੀ ਦੇ ਪੁਰਾਣੇ ਢੰਗ ਬਦਲ ਰਹੇ ਹਨ ।//

Pattern 3:-

Sub-ordinate Conjunction + Dependent Clause + Sub-ordinate Conjunction+ Independent Clause

These types of complex sentences are composed of two sub-ordinate conjunctions with one independent clause and one dependent clause.

/ /ਜੇ / / ਕਿਤੇ ਨਿਉਕਲੀ ਯੁੱਧ ਹੋ ਗਿਆ / /ਤਾਂ/ / ਮਾੜੇ ਯੁੱਧ ਨਾਲ ਹੀ ੨੫੦ ਸ਼ਹਿਰ ਜੰਗੀ ਕਹਿਰ ਦਾ ਨਿਸ਼ਾਨਾ ਬਣਨਾ ਲਾਜ਼ਮੀ ਹਨ ।/ /

Pattern 4:-

/ /Dependent Clause/ / +/ /, / /+/ / Independent Clause / /

These types of complex sentences in-dependent and dependent clauses are joined by comma.

/ / ਦਿੱਲੀ ਜਾਂਦਿਆਂ / / , / / ਮੈਂ ਰਾਹ ਵਿੱਚ ਜ਼ਾਖਲ ਉਤਰ ਗਿਆ ।/ /

Pattern 5:-

/ /Subject / / + / / Independent Clause / / + / / Predicate / / + / / Dependent Clause / /

In these types of complex sentences first comes the subject of the independent clauses. After this comes the dependent clause and in the end comes the predicate of the independent clause.

/ /ਇਹ ਪਾਰਟੀ / / ਜੋ / / ਮੁਕਾਬਲੇ ਦੀ ਸਰਕਾਰ ਦੀਆਂ ਫੜ੍ਹਾਂ ਮਾਰਦੀ ਸੀ,/ / ਸਾਰੀਆਂ ਸੀਟਾਂ ਤੇ ਹਾਰ ਗਈ । / /

Pattern 6:-

Independent Clause + Sub-ordinate Conjunction + Dependent Clause + Sub-ordinate Conjunction or (,) + Dependent Clause.

These types of complex sentences one dependent and two independent clauses occur. In the starting of the sentence there comes independent clause and after that come both the dependent clauses. First dependent clause is joined with sub-ordinate conjunction and the second one is joined with conjunction or comma.

/ / ਹੁਣ ਸਾਡੀ ਗੱਲਬਾਤ ਵਿੱਚੋਂ ਵੀ ਦਰਿਆ ਲੋਪ ਹੋ ਗਏ / / ਅਤੇ / / ਪਾਣੀ ਦਾ ਜ਼ਿਕਰ ਆਉਂਦਾ ਹੈ / / ਤਾਂ / / ਘਾਟ ਜਾਂ ਥੁੜ੍ਹ ਦੇ ਅਰਥਾਂ ਵਿੱਚ ਹੀ ਆਉਂਦਾ ਹੈ ।

Pattern 7:-

/ / Sub-ordinate Conjunction / / + / / Dependent Clause / / + / / Sub-ordinate Conjunction / / + / / Independent Clause/ / + / / Sub-ordinate Conjunction / / + / / Dependent Clause

These types of complex sentences contain two dependent clauses and one in-dependent clause. The independent clause lies in between two dependent clauses and each clause is starts with sub-ordinate conjunctions.

/ / ਜੇ / / ਕਿਸੇ ਦੇ ਘਰ ਪੁੱਤਰ ਜੰਮਿਆ / / ਤਾਂ / / ਉਹ ਜਰੂਰ ਕਿਸੇ ਗੁਰੂ, ਪੀਰ, ਅਵਤਾਰ, ਸੰਤ, ਮਹੰਤ ਦੀ ਬਖਸ਼ਸ਼ ਹੈ / / ਜਾਂ / / ਕਿਸੇ ਮੜ੍ਹੀ, ਮਸਾਣੀ, ਕਬਰ, ਮੱਠ ਜਾਂ ਮੂਰਤੀ ਦਾ ਫਲ ਹੈ । / /

Pattern 8:-

/ / Independent Clause / / + / / Sub-ordinate Conjunction / / + / / Independent Clause / /+ / / Sub-ordinate Conjunction / / + / /Dependent Clause

These types of complex sentences contain two in-dependent clauses and one dependent clause. Each clause starts with conjunction. In-dependent clause comes with co-ordinate conjunctions and dependent clauses come with sub-ordinate conjunctions.

/ / ਬਹੁਤ ਲੋਕ ਰੱਬ ਤੋਂ ਡਰ ਕੇ ਅਰਦਾਸ ਕਰਦੇ ਹਨ / / ਅਤੇ / / ਰੱਬ ਤੋਂ ਉਹ ਚੀਜ਼ਾਂ ਮੰਗਦੇ ਹਨ / / ਜਿਹੜੀਆਂ / / ਉਹਨਾਂ ਦੇ ਘਰ ਦੇ ਨੇੜਲੇ ਡਿਪੂ ਤੋਂ ਵੀ ਮਿਲ ਸਕਦੀਆਂ ਹਨ ।/ /

7. CHARACTERISTICS OF COMPLEX SENTENCES

1. The complex sentence contains at least one dependent and one independent clause. In case of multiple clauses it may also contain more than one dependent clause.
2. In complex sentence most of the clauses are joined by sub-ordinate conjunctions.
3. The dependent clause of the complex sentence may contain both finite and non-fine verb phrase.
4. There are three types of conjunctions used in the complex sentences. Simple, compound and complex.
5. In complex sentences are mainly classified in two categories i.e. predicate bound and non-predicate bound.

8. ALGOIRITHM USED FOR IDENTIFICATION OF COMPLEX SENTENCES

We identify the complex sentences on the basis of identification marker present in complex sentence.[9] The algorithm used is as follow:-

Step 1: Scan the whole sentence and check if it contains more than one verb. If it contains only one verb then it is simple otherwise it may be compound or complex and hence need further processing.

Step 2: Search for the presence of dependent clause. In Punjabi the dependent clause are broadly classified in two groups. These are Predicate bound clauses and other is Non predicate bound clauses. Further predicate bound clauses are of three types (Participial, Infinitival and Conjunctival). Non-predicate bound clauses are of two types (Sequential and Non-sequential).

Step 3 : All above type of clauses can be identified as shown in the following table:-

Name of the Clause	Identification Marker	Types	Identification using POS tag and others	Position in the sentence
Participial	Contains non-finite verb like ਟੱਪਦਿਆਂ, ਵੇਖਿਆ , ਕੀਤਿਆਂ	Two type:- Perfect & Imperfect	DIAN , NIAN	In the beginning of the sentence
Infinitival	Contains ਨੇ , ਨ , ਏ with root verb. E.g. ਕਰਨੇ, ਜਾਏ	Two type:- Simple & Imperfective	VBMA with ਨੇ , ਨ , ਏ	In the beginning of the sentence
Conjunctival	ਕੇ with root verb. E.g. ਖਾ ਕੇ , ਵੇਖ ਕੇ		PTUKE	In the beginning of the sentence
Sequential	Dependent clause		VBMA followed by	On second or last position

	starts with ਕਿ		ਕਿ	
Non-sequential	ਜਦੋਂ , ਜੇ, ਭਾਵੇਂ , ਕਿਉਂਕਿ etc	Two types:- Relative & Non-Relative	ਜਦੋਂ , ਜੇ, ਭਾਵੇਂ , ਕਿਉਂਕਿ	May be on first, second or last position

Step 4 : Using information of step 3 an application has been developed that identify and separate the complex sentences from Punjabi language corpus

9. RESULTS AND DISCUSSION

We tested our module on Punjabi corpus 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%

10. CONCLUSIONS AND FUTURE WORK

In this study, we proposed the initial implementation of identification of complex sentences on the basis of identification mark that helps in separating predicate bound
and non-predicate bound type complex sentences. In future we can use corpus based statistical approaches for identification of complex sentences. Also this research could be helpful in simplification of complex sentences in to simple sentences and hence can be used in machine translation.

11. REFERENCES

[1] Poornima C, Dhanalakshmi V, Anand Kumar M and Soman K P (2011) "Rule based Sentence Simplification for English to Tamil Machine Translation System", International Journal of Computer Applications (0975 – 8887)Volume 25– No.8.

[2] Zhemin Zhu, Delphine Bernhard and Iryna Gurevych 2010. "A Monolingual Tree based Translation Model for Sentence Simplification", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).

[3] Ani Thomas, M K Kowar, Sanjay Sharma. and H R Sharma. (2011) "Extracting Noun Phrases in Subject and Object Roles for Exploring Text Semantics", International Journal on Computer Science and Engineering (IJCSE) vol-3.

[4] Akshar Bharati, Dipti Misra Sharma, Sukhada (2009) Adapting Link Grammar Parser (LGP) to Paninian Framework Mapping of Parser Relations for Indian Languages', National Seminar on Computer Science and its Applications in Traditional Shastras (CSATS'09).

[5] Katsuhito Sudoh et al. 2010. "Divide and Translate:Improving Long Distance Reordering in Statistical Machine translation".

[6] Kamaljeet Kaur Batra and G S Lehal "Rule Based Machine Translation of Noun Phrases from Punjabi to English".

[7] Naushad UzZaman, Jeffrey P. Bigham and James F. Allen (2011) "Multimodal Summarization of Complex Sentences", IUI 2011, February pp. 13-16.

[8] Amitabha Mukerjee, Ankit Soni and Achla M Raina(2006) "Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora".

[9] Daraksha Parveen, Ratna Sanyal and Afreen Ansari (2011) "Clause Boundary Identification using Classifier and Clause Markers in Urdu Language".

[10] Mrs.M.Humera Khanam,Mr.Palli Suryachandra, Prof.K.V. MadhuMurthy "Dependency Parsing for Telugu", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,Vol.1, Issue 4, pp.1751-1754.

[11] "Introduction to Punjabi Grammer" InternetSource: http://punjabi.aglsoft.com/punjabi/learngrammar/