Identification and Separation of Simple, Compound and Complex Sentences in Punjabi Language

Chandni
Adhesh college of Engineering,
Faridkot

Rajneesh Narula
Lecturer, Adhesh Institute of
Engineering and Technology
Faridkot,

Sanjeev Kumar Sharma
Assistant Professor
DAV UniversityJalandhar

Abstract:

Sentence identification is one of the basic applications of the Natural Language Processing. In this paper we have explored the different types of sentences present in Punjabi language. Some detail of the internal structure of these sentences has also been discussed. We have also proposed an algorithm for identification of simple, compound and complex sentences. The work done on the compound and complex sentences in various Indian languages have also been discussed.

Keywords:

PunjabiSentences, Sentence Identification, Simple, compound and complex sentences.

Introduction:

Natural language processing is a very young discipline in Punjabi. Therefore, there is a lack of basic resources and tools for processing the Punjabi language. Sentence Identification is one of the activities performed in a typical word processing application. Sentence Identification means to ensure that a given piece of text follows the grammar rules and structure of specific type of sentence of language in which it is written. Sentence Identification systems for some of foreign languages have been directly or in- directly available but at present, no such system is available for any of the Indian languages. With the computers being widely used for day-to-day tasks of word processing,grammar checking, summarization etc. need for sentence identification is being felt earnestly.

Introduction to Punjabi sentences:

Punjabi (or Panjabi) language is a member of the Indo-Aryan family of languages, also known as Indic languages.The sentences are composed of clauses and the simplest form of a sentence has only one independent clause. The structure of Punjabi language's sentences is SOV (Subject Object Verb) order unlike English that follows SVO order. The subject occurs first followed by the object and then the verb. e.g.

Sentence 1: ਮੁੰਡਾ ਸੇਬ ਖਾਂਦਾ ਹੈ।muṇ ḍ ā sēb khāndā hai .

In the above sentence 1, ਮੁੰਡਾmuṇ ḍ ā'boy' is subject, ਸੇਬsēb 'apple' is object, and ਖਾਂਦਾ ਹੈkhāndāhai 'eats' is verb. The nominal phrases behave as subject and object in a clause or sentence, and the verb phrases form the verb part of that clause or sentence. Based on the structure, the sentences can be classified into three categories:

Simple Sentence:A sentence that contains only one clause and that clause is independent clause is called simple sentence.E.g.

Sentence 2: ਮੁੰਡਾ ਗੀਤ ਗਾ ਰਿਹਾ ਹੈ।muṇ ḍ ā gīt gārihā hai .

In above sentences there is only one independent clause and hence it is a simple sentence.

Compound Sentence: A sentence that contains two or more than two independent clauses joined by coordinate conjunctionslike ਜਾਂjāṃ 'or' ,ਅਤੇ/ਤੇatē/tē 'and'are called compound sentences e.g.:

Sentence 3: ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ ਤੇ ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ।

mīṃh pai rihā sī tē lōk bhijj rahē san .

Sentence 4: ਹੁਣ ਮੈਂ ਸ਼ਹਿਰ ਜਾ ਰਹੀ ਹਾਂ ਫੇਰ ਤੂੰ ਚਲਾ ਜਾਈਂ।

huṇ maiṃ sa◌̣hir jā rahī hāṃ phēr tūṃ calā jāīṃ .

In above sentence 3 contains two independent clauses i.e. ਮੀਂਹ ਪੈਰਿਹਾ ਸੀ and ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ joined by conjunction ਤੇ. Similarly in sentence 4 two independent clauses are ਹੁਣਮੈਂਸ਼ਹਿਰਜਾਰਹੀਹਾਂ and ਤੂੰਚਲਾਜਾਈਂ joined by conjunction ਫੇਰ [17].

Complex Sentence: A sentence that contains at least one dependent clause with independent clause is called complex sentence. These two clauses are combined by using subordinate conjunctionsਤਾਂtāṃ‘then’, ਜੇjēif’etc, e.g.:

Sentence 5: ਜੇਤੂੰਸਾਡੇਨਾਲ਼ਜਾਣਾਹੈਤਾਂਆਜਾ।

jētūṃsāḍ ēnāla◌̣jāṇ āhaitāṃ ā jā.

If you want to come with us then come on.

In above sentence 5ਜੇjē ‘if and ਤਾਂtāṃ ‘then’ are sub-ordinate conjunctions. They occurs in pair i.e. one part lies in the beginning of dependent clause and other part lies in the beginning of independent clause. In above sentence ਜੇਤੂੰਸਾਡੇਨਾਲ਼ਜਾਣਾਹੈis the dependent clause starting with subordinate conjunction ਜੇ and the other part ਆਜਾis the independent clause joined with dependent clause with the second part of sub-ordinate conjunction (ਜੇ -ਤਾਂ)i.e. ਤਾ [17].

Previous work done:

NaushadUzZaman , Jeffrey P. Bigham and James F. Allen (2011) [2] 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.

Amitabha Mukerjee, AnkitSoni and Achla M Raina (2005) [3] have developed Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. They constructed the first corpus-based lexicon of Complex Predicates in Hindi based on projecting POS tags across parallel English-Hindi corpora. The CP types considered include adjective-verb (AV), noun-verb (NV), adverb-verb (Adv-V), and verb-verb (VV) composites. A verb in English is projected onto a multi-word sequence in Hindi where CPs are hypothesized.The resulting database lists usage instances of 1439 CPs in 4400 sentences. While such approaches sometimes leave out some CPs, the ones that are named are seen to be quite robust. As a result, this appears to be a good first approach for identifying the majority of CPs along with usage data. Since the approach involves minimal linguistic analysis, it is easily extendable to other languages which exhibit similar CP constructs, provided the availability of a POS lexicon.

DarakshaParveen, RatnaSanyal and Afreen Ansari (2009) [4]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 character to detect the type of sub-ordinate clause that 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 indicate 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.

Lucia Specia, Research Group in Computation Linguistics University of Wolverhampton, UK has developed Translating from Complex to Simplified Sentences and presented new approach for text simplification that was based on the framework of Statistical Machine Translation. Statistical Machine Translation framework was used to learn how to translate from complex to simplified sentences. In this research, Corpus of natural simplifications was used to train the translation system. The experiments had shown that the framework can appropriately simplify certain phenomena, those related to lexical operations, lexical simplifications and simple rewriting. Text simplifications had been used for improving the accuracy of the natural language processing tasks. Some additional feature like part-of-speech tags had been used as factors in standard phrase-based systems like Moses.

LEFFA, Vilson Jose (2008) [5] has developed Clause Processing in Complex Sentences. They proposed and test an algorithm for the segmentation of complex sentences into clauses. The algorithm is built later the parts of speech for each lexical item are assigned. The algorithm was tested using a machine translation system, which included an English/Portuguese dictionary, a part of speech tagging system and the ability to introduce rules, including the ones required by the algorithm. The consequences showed that out of 1659 clauses, randomly taken from a 10,000,000-word corpus, more than 98% were correctly segmented and 95%correctly classified into nouns or adverbs. The major problems found in segmenting and classifying the clauses included conjunction ambiguity, the sharing of the same subject by different clauses and verbs that belonged to more than one sub categorization.

Sander Wubben, Antal van den Bosch, EmielKrahmer(2012)[19] have developed Sentence Simplification by Monolingual Machine Translation. They described a method for simplifying sentences using Phrase Based Machine Translation, grew with a re-ranking heuristic based on dissimilarity, and trained on a monolingual parallel corpus.They compare their system to a word-substitution baseline and two state-of-the-art systems, all trained and tested on matched sentences from the English part of Wikipedia and Simple Wikipedia. Human test subjects estimate the output of the different systems. Examining the judgments shows that by relatively careful phrase based paraphrasing our model achieves similar simplification results to state-of-the-art systems, while generating better formed output. They also argued that text readability metrics such as the Flesch-Kincaid grade level should be used with caution when evaluating the output of simplification systems.

Mandeep Singh Gill,Gurpreet Singh Lehal (2008) [10] have developed a Grammar Checking System for Punjabi.They provided description about the grammar checking system developed for detecting various grammatical errors in Punjabi texts. This system applies a full-form lexicon for morphological analysis, and uses rule-based approaches for part-of-speech tagging and phrase chunking. The system adopts a novel approach of performing agreement checks at phrase and clause levels using the grammatical information exhibited by POS tags in the form of feature value pairs. The system can observe and propose rectifications for a number of grammatical errors, resulting from the deficiency of agreement, order of words in several phrases etc, in literary style Punjabi texts. This grammar checking system is the first such system reported for Indian languages.

Navneet kaur, Sanjeev kumar Sharma and KamaldeepGarg (2013) [11] proposed a system for identification and separation of complex sentences from Punjabi language. They explained different types of complex sentences and used the identification marker to separate out the complex sentences. They further divided the complex sentences in to two categories i.e. predicate bound sentences and non-predicate bound sentences. Their system takes the Punjabi corpus as input and separate out the predicate bound and non-predicate bound sentences. They obtained an accuracy of 83.5 for predicate bound sentences and 80.5% for non-predicate bound sentences.

Sanjeev Kumar Sharma and Dr G S Lehal (2014)[12] in their research paper proposed a method for identification of compound sentences. They explain different type of compound sentences and their patterns. They proposed an algorithm for identification of compound sentences on the basis of co-ordinate conjunction and using the phrases of the sentences. They used a HMM based POS tagger and a rule based phrase chunker for preprocessing and identification. They obtained an accuracy of 86.5%.

Our approach:

Punjabi language like other Indian languages is very inflective in nature and thus some specific feature of this language can be used for identification of different types of sentences. E.g. for identification of complex sentences we need to identify the dependent clause. A dependent clause can be identified by using special feature of dependent clause like presence of finite verb in the dependent clause, presence of a specific postposition after the root form of the verb etc. similarly for compound sentences we need to identify the presence of multiple independent clauses in a sentence. These multiple independent clauses can be identified with the presence of more than one verb in the sentence.If there is only one independent clause then this will be a simple sentence. We extract these featured from a pre-annotated corpus of compound and complex sentences. The feature used has been shown in the following table:

Table 1
Feature used for identification of compound and complex sentences

Sr. No	Type of sentence	Name of the feature	Example	identification
1.	Complex	Presence of non-finite verb	ਟੱਪਦਿਆਂ, ਵੇਖਿਆ , ਕੀਤਿਆਂ	Can be identified by presence of suffix DIAN and NIAN in POS tag.
2.	Complex	Contains ਨੇ ,ਨ , ਣੇ with root verb.	ਕਰਨੇ, ਜਾਣੇ	Can be identified by presence of ਨੇ , ਨ , ਣੇ as suffix with the root word and having tag VBMA
3.	Complex	Presence of ਕੇ with root verb.	ਖਾਕੇ , ਵੇਖਕੇ	Presence of ਕੇ or its tag PTUKE after main verb
4.	compound	Presence of more than one main verb in the sentence		Can be identified by counting the main verbs

Proposed Algorithm:

All the three types of sentences i.e. simple, compound and complex can be identified on the basis of the clauses present in them. We proposed the following algorithm for identification and separation of simple compound and complex sentences:

Step 1: Input Punjabi Sentence in Unicode
Step 2: Apply morph and Part of speech Tagger
Step 3: COUNT the no of main verbs present in the sentence.
Step 4: if the COUNT=1 goto Step 8
Step 5: Check for the Presence of dependent clause by using the feature mention in table 1. If present then goto Step 7.
Step 6: separate the sentence as compound sentence and Exit.
Step 7: separate the sentence as complex sentence and Exit.
Step 8: separate the sentence as Simple sentence and Exit.

Potential Use:

Following are some of the application areas where this complex sentence identification system can be used. It is obvious that the key application of this system is in word processing environment. The system as a whole and its subsystems will find numerous applications in natural language processing of Punjabi. Following are some of the application areas of this system as a whole or its subsystems:

- It could be used with various information processing systems for Punjabi, where the input needs to be corrected before processing. For instance, machine translation systems translating texts from Punjabi to other languages may use this system for simplification of sentences of the input Punjabi text.
- This system is very helpful in summarization of the sentences. All the complex sentences can be separated out.
- This system as a whole could be used as a sub part for the development of Grammar checker for compound and complex sentences.
- Next language learners of Punjabi could use this system as a writing aid to learn grammatical categories operating in Punjabi sentences, and thus improve their writings by learning from their mistakes.

Experimental Evaluation

The accuracy of Natural Language product is generally measured in terms precision and recall. Precision is the percentage of correctly identified sentences. And recall is If A is the number of correctly identified sentences and B is the number of sentences that were not identified by our system then

Recall = A / (A+B)
Similarly if C is the no of correctly Identified sentences and D is the number of incorrectly identified sentences then
Precision= C / (C+D)
For evaluation of the proposed sentence identifier, a corpus having texts from different online resources i.e. Punjabi websites were used. The outcome was manually evaluated through a linguistic expert to mark the correct and incorrect sentences. 1100 sentences were collected randomly.The result obtained have been given in Table 2.

Table 2
Experimental Result

Total No of Sentences	1100
Number of simple sentences	711
Number of Correctly identified simple sentences	695
Number of in-correctly identified simple sentences	7
Number of compound sentences	74
Number of correctly identified compound sentences	66
Number of in-correctly identified compound sentences	7
Number of complex sentences	325
Correctly identified complex sentences	309
Number of in-correctly identified complex sentences	11

For simple sentence:
Precision = 695/ (695+7)
0.99
Recall = 695/ (695+16)
0.97
For compound sentence:
Precision = 66/ (66+7)
0.90
Recall = 66/ (66+8)
0.89
For complex sentence:
Precision = 309/ (309+11)
0.96
Recall = 309/ (309+16)
0.95

Conclusion:

In this paper we have proposed an algorithm for identification of simple compound and complex sentences. Our proposed system gives accuracy more than 90% in terms of precision and recall. This can be further improved by using some other statistical techniques like support vector machine and conditional random field etc. This algorithm can also be implemented on other Indian languages havingsentence structure similar to Punjabi.

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

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

[4]. DarakshaParveen, RatnaSanyal and Afreen Ansari (2011) "Clause Boundary Identification using Classifier and Clause Markers in Urdu Language".

[5]. LEFFA, Vilson Jose (2008), "clause processing in complex sentences", Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Espanha: 1998. v. 2, p. 937-943.

[6]. Jonah Lin (2005)"Syntactic structures of complex sentences in Mandarin Chinese",National Tsing Hua University.

[7]. Deepthi Chidambaram (2005) "Processing Complex Sentences for Information Extraction" , Arizona State University.

[8]. David Vickrey,DaphneKoller (2008) **"**Sentence Simplification for Semantic Role Labeling" Stanford University Stanford, CA 94305-9010.

[9]. SiddharthaJonnalagadda,LuisTari, JorgHakenberg,ChittaBaral, Graciela Gonzalez(2009) "Towards Effective Sentence Simplification forAutomatic Processing of Biomedical Text",Proceedings of NAACL HLT Short Papers, Association for Computational Linguistics.

[10]. Mandeep Singh Gill, Gurpreet Singh Lehal (2008) "Grammar Checking System for Punjabi" Coling 2008:companion volume Posters and Demonstrations pages 149–152 Manchester.

[11]. Navneet kaur, Sanjeev kumar Sharma and KamaldeepGarg (2013) "Identification and separation of complex sentences from Punjabi language" International Journal of Computer Applications (0975 – 8887) Volume 69– No.13, May 2013 pp 21-24.

[12]. Sanjeev Kumar Sharma and Dr G S Lehal (2014) "Identification of compound sentences in Punjabi language" Research Cell: An International Journal of Engineering Sciences, Inaugural Issue 2010 ISSN: 2229-6913

[13]. Alam, Md. Jahangir, NaushadUzZaman, and MumitKhan(2006), "N-gram based Statistical Grammar Checker for Bangla and English",InProc. ofninthInternational Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh.

[14]. Chander, Duni. 1964. Punjabi Bhasha da Viakaran(Punjabi). Punjab University Publication Bureau, Chandigarh, India.

[15]. Gill, Harjeet S. and Henry A. Gleason, Jr. (1986), "A Reference Grammar of Punjabi", Publication Bureau, Punjabi University, Patiala, India.

[16]. Puar, Joginder S. (1990), "The Punjabi verb form andfunction",Publication Bureau, Punjabi University, Patiala, India.

[17]. "Introduction to Punjabi Grammar" Internet source http://punjabi.aglsoft.com/punjabi/learngrammar/

[18]. Sander Wubben, Antal van den Bosch,EmielKrahme (2012),"sentence simplification by monolingual machine translation" acl pp 1015-1024

[19]. Lucia Specia. Translating from Complex to Simplified Sentences. Lecture Notes in Computer Science, 6001:30–39, Springer-Verlag, 2010.