Using Sentence Simplification to Generate Paraphrase for Low Resource Punjabi Language

[1]Ravinder Mohan Jindal,[2]Leekha Jindal,[3]Sanjeev Kumar Sharma
[1]Research Scholar, [2]Research Scholar, [3]Associate Professor
[1,2] SBBS University, Jalandhar
[3]DAV University, Jalandhar
[1]rmjindal2002@gmail.com, [2]leekhajindal@gmail.com, [3]sanju3916@rediffmail.com

ABSTRACT

The field of natural language processing is growing in computer science, and generating paraphrases is a difficult task, especially for languages like Hindi, Punjabi, and Urdu, which are morphologically rich and have limited resources. This research article focuses on generating paraphrases for Punjabi, a morphologically rich Indian language, using a sentence simplification approach. The author employed several sentence simplification algorithms to simplify long Punjabi sentences and used antonym-synonym replacement to generate the paraphrases. The sentence simplification component of the system achieved a precision of 100%, recall of 95%, and an f-measure of 97.43% when tested with a set of data. The developed system's performance was analyzed using various complexity measurement parameters, and it was observed that a combination of lexical and syntactic simplifications yielded the best results.

Keywords: NLP, Punjabi language Processing, Paraphrasing, Syntactic simplification, Lexical simplification

INTRODUCTION

Paraphrasing is the process of restating a text or idea using different words while preserving the original meaning. It involves understanding the main points of the original text and expressing them in a new way, usually with a different sentence structure and choice of words. Paraphrasing is an essential skill in academic writing and research as it allows writers to integrate ideas from other sources into their work while avoiding plagiarism. It also helps to improve comprehension and clarity by presenting complex ideas in simpler terms. However, it is important to note that paraphrasing should not alter the original meaning of the text, and proper citation should be given to the source material. Effective paraphrasing requires careful reading, analysis, and interpretation of the original text, and the ability to express its ideas in a clear and concise manner.

Punjabi is an Indo-Aryan language spoken by over 100 million people, primarily in the Punjab region of India and Pakistan. It is the official language of the Indian state of Punjab and the second most widely spoken language in Pakistan. Punjabi is also spoken by Punjabi communities around the world, including in Canada, the United Kingdom, and the United States.

Punjabi is a tonal language with three tones: high, mid, and low. It is a richly inflected language with ten noun cases, which are used to indicate the relationship between the noun and the verb. Punjabi also has a complex system of verbs, which are conjugated according to tense, aspect, mood, and person. The Punjabi script is based on the Gurmukhi script, which was developed in the 16th century by the Sikh guru, Guru Angad Dev. The script has 35 letters, including three vowels and 32 consonants. It is written from left to right and is phonetic, meaning that each letter represents a distinct sound. Punjabi has a rich literary tradition, with works dating back to the 11th century. The language has produced many notable poets, including Bulleh Shah, Waris Shah, and Amrita Pritam. Punjabi literature includes a variety of genres, including poetry, prose, and drama. Punjabi is a language that is constantly evolving, with new words and phrases being added to the lexicon. The language has also been influenced by other languages, such as Urdu and Hindi, due to their close proximity and cultural exchange.

Despite its rich cultural heritage, Punjabi is facing challenges in the modern era. The language is struggling to keep up with the rapid pace of technological and economic development, and many young Punjabis are turning to English and other languages for education and job opportunities. Efforts are being made to promote and preserve the Punjabi language, including the establishment of Punjabi language academies and the introduction of Punjabi language courses in schools and universities. However, more needs to be done to ensure that the language continues to thrive and evolve in the 21st century.

LITERATURE REVIEW (EXISTING WORK):

Paraphrasing is an essential skill in academic writing, research, and communication. The ability to restate a text or idea using different words while preserving the original meaning is crucial for avoiding plagiarism, improving comprehension, and presenting complex ideas in simpler terms. Over the years, various approaches have been developed for paraphrasing, each with its own strengths and weaknesses.One of the earliest approaches to paraphrasing is the substitution method. This approach involves replacing words or phrases in the original text with synonyms or near-synonyms while preserving the sentence structure and grammar. While this approach is straightforward and can be effective for simple sentences, it can be problematic for more complex texts, where the choice of synonyms may alter the meaning of the original text or result in awkward phrasing.Another approach to paraphrasing is the sentence restructuring method. This approach involves changing the sentence structure while preserving the meaning of the original text. This can be achieved by changing the order of words, using passive voice instead of active voice, or changing the tense or aspect of the verb. While this approach can be effective for improving clarity and readability, it can also result in a loss of precision and may not be suitable for technical or scientific writing.The sentence fusion method is another approach to paraphrasing that involves combining two or more sentences into a single sentence while preserving the meaning of the original text. This approach can be effective for simplifying complex ideas and improving flow and coherence. However, it can also result in longer sentences and may not be suitable for all types of writing.The sentence simplification method is a more recent approach to paraphrasing that involves simplifying complex sentences by removing or replacing certain elements. This can include removing redundant or irrelevant information, replacing complex words with simpler ones, or breaking down complex sentences into smaller, more manageable parts. This approach has shown promise in generating paraphrases for morphologically rich and low-resource languages, such as Hindi, Punjabi, and Urdu.

While each of these approaches has its own strengths and weaknesses, it is important to note that effective paraphrasing requires careful reading, analysis, and interpretation of the original text, and the ability to express its ideas in a clear and concise manner. In addition, proper citation should be given to the source material to avoid plagiarism. As technology continues to evolve, new tools and techniques for paraphrasing are being developed, including machine learning algorithms and natural language processing techniques.

METHODOLOGY USED

The author utilized a two-step approach to produce paraphrases of Punjabi language, which involved employing syntactic paraphrase or sentence simplification, followed by lexical paraphrasing or synonym-antonym replacement. During the sentence simplification process, the author broke down compound and complex sentences into simpler ones. Afterward, by replacing antonyms and synonyms, the author created paraphrases from these simplified sentences.

Flow chart to implement all above steps are shown in figure 1

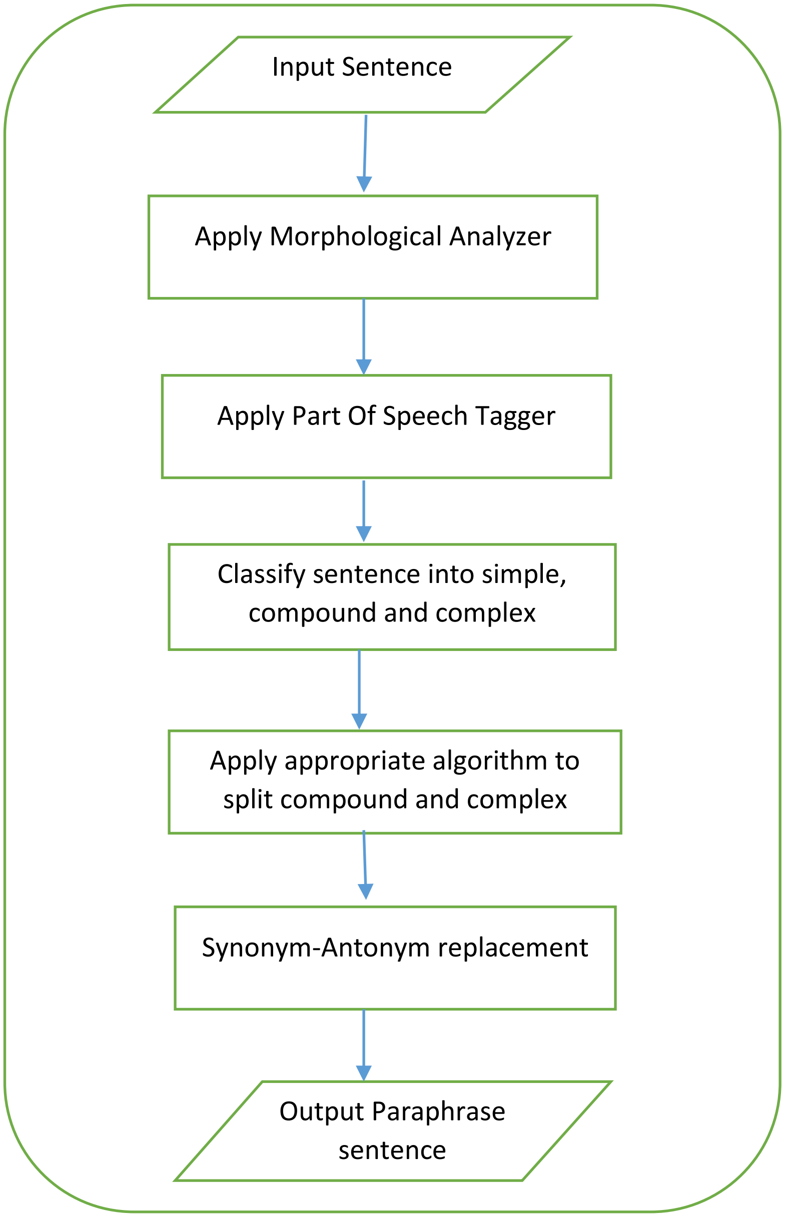
Input Sentence
Apply Morphological Analyzer
Apply Part Of Speech Tagger
Classify sentence into simple, compound and complex
Apply appropriate algorithm to split compound and complex
Synonym-Antonym replacement
Output Paraphrase sentence

Figure 1: Flow diagram of various steps used for paraphrasing

The procedure outlined in Figure 1 is executed in a sequential manner. The entire paraphrasing process is divided into four distinct phases, namely sentence classification, sentence pre-processing, sentence simplification, and sentence paraphrasing. Each phase occurs one after the other, and the output of the preceding phase is utilized as the input for the subsequent phase.

During the sentence classification phase, the input sentences are classified into simple, compound, and complex categories. The sentence pre-processing phase involves utilizing spell checkers and grammar checkers to identify and correct any lexical or syntactic errors. In the sentence simplification phase, a variety of simplification algorithms are employed to divide compound and complex sentences into simpler ones. Further details on each of these phases, along with examples, are elaborated on below.

Phase1 (Sentence classification)

This marks the initial stage of our research project, wherein the first step involves breaking down a paragraph or corpus into individual sentences. Subsequently, these sentences are categorized into several sentence types, including simple, compound, complex, and compound-complex sentences. To achieve this classification, both the structure of the sentence and morphological characteristics of the Punjabi language are taken into consideration. In addition to these factors, pattern matching techniques are also employed, and the outcomes from both methods are compared to yield the final classification.

A. Phase 2 (Pre-processing)

During this phase, the classified sentences obtained in phase 1 are subjected to pre-processing. The primary objective of pre-processing is to eliminate any spelling and grammar errors in the input text. This involves conducting both spell checking and grammar checking of the sentences. Spell checking is necessary for identifying any incorrectly spelled words, as it would not be feasible to retrieve synonyms and antonyms for words with spelling errors from the databases. Additionally, spell checking serves as a fundamental requirement for grammar checking. Grammar checking is conducted to identify any syntactic errors in the sentence since it can be challenging to break down a sentence that contains syntactic errors. To conduct both spell and grammar checking, existing tools developed by Punjabi University Patiala are utilized.

B. Phase 3 (Sentence Simplification)

This phase constitutes the primary stage of our research endeavor, wherein the objective is to break down large and intricate sentences into smaller, simpler ones. To achieve this, several algorithms have been designed to convert compound and complex sentences into simple ones. Furthermore, this phase encompasses the identification of clauses, including dependent and independent clauses, along with the use of separation algorithms. The algorithms implemented in this phase vary based on the structure of the sentence. In Punjabi language, the structure of complex sentences can vary depending on the number and position of dependent clauses, which necessitates the development of a distinct algorithm for each type of complex sentence.

C. Phase 4 (Sentence Paraphrasing)

This phase marks the final stage of the paraphrasing process, where the classified, pre-processed, and simplified sentences are transformed into their corresponding paraphrases. As illustrated in

Figure 6, each sentence slated for paraphrasing is subject to reframing rules, wherever applicable, to alter its syntax. These reframing rules have been designed and stored in a dedicated database. Subsequently, where possible, synonym and antonym replacements are executed to create two alternative sentences. Since there is no existing database for synonym and antonym replacement in Punjabi language, a new database has been created from scratch. Finally, the two alternative sentences created by replacing synonyms and antonyms are combined to form a new paraphrased sentence.

RESULT AND DISCUSSIONS:

The ILCI corpus was utilized to test the performance of the developed system. This corpus consists of texts from diverse domains and varies in length. The corpus encompasses data from a variety of fields, including agriculture, entertainment, tourism, health, and news. The specifics of the corpus, such as the amount of data it contains, are provided in Table 1.

Table 1: details of the ILCI corpus used

Sr. No.	Type of Corpus/ Domain of the corpus	Number of files	Total Number of Sentences in the file (approximated in thousand (k))	Number of words (approximated in thousand (k))
1	Agriculture corpora (AC)	20	22	254
2	Entertainment corpora (EC)	20	23	276
3	Tourism corpora (TC)	19	22	255
4	Health corpora (HC)	25	31	372
5	News corpora sports news (NC)	12	15	180

Table 2: Performance of sentence classification module

Type of Corpus/ Domain of the corpus	(Overall sentences given to system in 000) (x)	Correctly separated (in simple, compound and complex in 000) by system (y)	Incorrectly separated by system (in 000) (z)	Precision $\frac{y+z}{x}$ X 100	Recall $\frac{y}{x}$ X 100	F-measure $\frac{Preciion\ X\ Recall}{Precision+Recall}$ X 2
Agriculture	22	21.6	0.4	100	98.2	99.09

corpora (AC)						
Entertainment corpora (EC)	23	21.06	1.94	100	91.6	95.62
Tourism corpora (TC)	22	21.03	0.97	100	95.6	97.75
Health corpora (HC)	31	30.13	0.87	100	97.2	98.58
News corpora sports news (NC)	15	30.86	1.14	100	92.4	96.05
Overall	113	124.68	5.32	100	95	97.43

CONCLUSION AND FUTURE SCOPE:

In this research paper, the author introduces a new methodology for producing paraphrases of Punjabi sentences. The approach entails utilizing syntactic and lexical paraphrasing. The author used sentence simplification for syntactic paraphrasing, breaking down long sentences (compound, complex and compound-complex sentences) into simple sentences using the appropriate algorithm. For lexical paraphrasing, the author employed synonym-antonym replacement. The results obtained from the analysis are presented in table 8, table 9, and table 10. The author concluded that the combination of syntactic and lexical paraphrasing is more effective than relying on only one method. The author conducted a literature review, revealing that no such work had been done on any Indian language except Hindi language. For Hindi language, the author only used lexical paraphrasing via synonym-antonym replacement. Future studies may employ deep learning and machine learning methods for implementing this paraphrasing task. Additionally, the proposed approach may be applied to produce paraphrases for other Indian and foreign languages.

REFERENCES

[1]. Lehal, G. S. (2007). Design and implementation of Punjabi spell checker. International Journal of Systemics, Cybernetics and Informatics, 3(8), 70-75.

[2]. Gill, M. S., Lehal, G. S., & Joshi, S. S. (2008). A punjabi grammar checker. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.

[3]. Gill, M. S., Lehal, G. S., & Joshi, S. S. (2009). Part of speech tagging for grammar checking of Punjabi. The Linguistic Journal, 4(1), 6-21.

[4]. Singh, D. M. (2010). A Punjabi Morphological Analyzer and Generator. Advanced Centre for Technical Development of Punjabi Language, Literature and Culture, Punjabi University.

[5]. Lehal, G. S. (2009). A Gurmukhi to Shahmukhi transliteration system. In proceedings of ICON-2009: 7th international conference on Natural Language Processing (pp. 167-173).

[6]. Goyal, V., &Lehal, G. S. (2009). Hindi-Punjabi Machine Transliteration System (For Machine Translation System). George Ronchi Foundation Journal, Italy, 64(1), 2009.

[7]. Josan, G. S., &Lehal, G. S. (2008, August). A Punjabi to Hindi machine translation system. In 22nd International Conference on on Computational Linguistics: Demonstration Papers (pp. 157-160). Association for Computational Linguistics.

[8]. Lehal, G. S., & Singh, C. (2000, September). A Gurmukhi script recognition system. In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000 (Vol. 2, pp. 557-560). IEEE.

[9]. Gupta, V., &Lehal, G. S. (2012, December). Automatic Punjabi text extractive summarization system. In Proceedings of COLING 2012: Demonstration Papers (pp. 191-198).

[10]. Kevin Knight and Daniel Marcu. 2000. Statisticsbased summarization-step one: Sentence compression. In Proceedings of AAAI-IAAI.

[11]. Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In Proceedings of COLING.

[12]. Katja Filippova and Michael Strube. 2008. Dependency tree based sentence compression. In Proceedings of INLG

[13]. Emily Pitler. 2010. Methods for sentence compression. Technical report, University of Pennsylvania.

[14]. Katja Filippova, Enrique Alfonseca, Carlos Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In Proceedings of EMNLP.

[15]. Kristina Toutanova, Chris Brockett, Ke M. Tran, and SaleemaAmershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In Proceedings of EMNLP.

[16]. Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient creation of an accurate sentence fusion corpus. In Proceedings of NAACL-HLT.

[17]. Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In Proceedings of COLING.

[18]. Dras, Mark. 1997a. Representing Paraphrases Using S-TAGs. Proceedings of the 35th Meeting of the Association for Computational Linguistics, 516-518.

[19]. Mark Dras. 1999. Tree adjoining grammar and the reluctant paraphrasing of text. Ph.D. thesis, Macquarie University, Australia

[20]. Regina Barzilay and Kathleen R McKeown. 2001. Extracting paraphrases from a parallel corpus. In Proceedings of ACL.

[21]. Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In Proceedings of ACL.

[22]. Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: Data and evaluation. In Proceedings of INLG.

[23]. Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In Proceedings of EACL.

[24]. AdvaithSiddharthan. 2010. Complex lexico-syntactic reformulation of sentences using typed dependency representations. In Proceedings of INLG.