International Journal of Mechanical Engineering

Punjabi-English Mixed Language Queries for Web Search

Harjit Singh[1,a)]

Author Affiliations

[1] *APS Neighbourhood Campus, Punjabi University, Patiala, India*

Abstract. Globalization is affecting the cultures as well as local languages. People became habitual to use English words in their non-English communication languages. Punjabi community is not away from this effect. People who are not well-versed in English, try to search the web in Punjabi-English mixed language queries. The search engines are not capable to handle such queries properly. This paper presents a methodology to handle these queries. The objective is to generate an equivalent English query from mixed language query. The process uses various modules to successfully achieve the objective. The native Punjabi words are translated using a bi-lingual dictionary. The Gurmukhi scripted English words are converted to Roman using a Gurmukhi-Roman List. The Punjabi nouns are transliterated using a rule based technique. The methodology is tested using a set of mixed language web queries and results are promising.

Keywords. Natural Language Processing, Information Retrieval, Bilingual Web Queries, Mixed Language Web Search

INTRODUCTION

This is the era of Internet and almost everybody is using web search to get the required information. The information on internet is being made available in multiple languages [1,2]. Globalization resulted in cultural changes and the people are using mixed languages to communicate with each other. Mostly, the local language is mixed with English by the people whose language is non-English. In almost every sentence, English words are frequently mixed with non-English native language during communication. It happens in oral as well as written communication such as messaging, blogging etc. The Punjabi speaking community is also being influenced by this cultural change. While speaking or messaging, the Punjabi speaking people are using Punjabi-English mixed language sentences. Even the people, who are not well-versed in English, try to use common English words in their daily routine. These people are not away from web searching. To clear their doubts, they often try to search the web. These people try to use Punjabi-English mixed language web queries. Such queries are not properly handled by the search engines. This paper presents a way to handle Punjabi-English mixed language web queries.

Punjabi Language

Punjabi is a language spoken by the people of Punjab state of India and related Punjabi community in other countries such as Pakistan, Canada, UK, US, UAE etc. [3, 4]. The total Punjabi speaking population counts to more than 100 million world-wide [5]. Punjabi community uses two types of scripts to write i.e. Gurmukhi and Shahmukhi [6]. This paper deals with Gurmukhi scripted Punjabi language.

LITERATURE REVIEW

Rozsa et al.[7] published a paper to report the results of a study of the behavior of a group of web searchers when English is used as a foreign language. The non-English native language users have to use English queries to get proper information because quality information may not be available in their native language. The authors used qualitative research methods for their study. Hengyi [8] conducted semi-structured interviews and analyzed the query log to study the patterns and strategies used by users to reformulate the Chinese-English queries. The study was performed on web searchers who used Chinese and English mixed language for their communication and web search. The information is very important to improve the performance of search engines and get more relevant results.

Xiaoyi and Mark [9] presented a method to search the web for bilingual text. The availability of parallel corpus is very limited even it is a useful resource for multilingual information retrieval. So the researchers presented a method to automatically gather multilingual text from the web. They tested their system on German-English bilingual text to prove that it was successful. Gao et al. [10] proposed a method to improve the web search by making use of the bilingual text. They performed their experiments on queries stored in query logs. They used Chinese queries and their equivalent English queries from query logs based on similar search interests. For each pair they performed the ranking of documents. Then, from this bilingual ranking, they generated monolingual ranking for Chinese.

Kwok [11] made use of Chinese-English wordlist for cross language IR using English-Chinese queries. It was proved that the wordlist can be used as word and phrase dictionary and it is a better option than using English-Chinese version. Zhang et al. [12] analyzed the search engines support for multiple languages. Since the information on internet is being made available in multiple languages, it is very important for search engines to support these languages. The authors analyzed the existing search engines and identified those having multi-lingual support. They concluded that EZ2Find, Google and

OnlineLink are the best in multilingual features. A workshop [13] was conducted to promote the research on internet search using non-English languages. The motive of workshop was to discuss the limitations of search engines to respond to non-English web queries. It was concluded that multi-lingual support by search engines will make them more effective and web searchers behavior should be examined for search engine improvements.

Lewandowski [14] tested the popular search engines for their capability to differentiate among English and German text. They used common words of both these languages to submit 50 web queries to search engines. They found that there were problems when MSN and Google were used with foreign language restriction mode. Gao et al. [15] provided a way to make use of search engine features (Domain Taxonomy, Page Rank etc.) for Chinese web queries. These features were basically developed for English web text. The authors analyzed the query logs for both English and Chinese languages to identify Chinese-English pair of queries. They used this information to improve ranking of Chinese web pages.

Munye and Atnafu [16] developed a bilingual search engine for Amharic and English. Popular search engines are unable to handle non-English queries, so this search engine was designed to handle Amharic queries and retrieve information in Amharic and English. The authors used NLP techniques for query preprocessing and a bidirectional translator for query translation. Two separate search engines for Amharic and English were used for experimentation and the results were promising.

METHODOLOGY

The input Punjabi-English mixed language web query is processed by several modules to reformulate the query. Most of the people, who are not comfortable to speak English, prefer to write English words in local language script, which, in this case is Gurmukhi. For example, Punjabi speaking person spells the word “Call” in Gurmukhi as “ਕਾਲ”, Rest as “ਰੈਸਟ”, Restaurant as “ਰੈਸਟੋਰੈਂਟ”, Admission as “ਅਡਮਿਸ਼ਨ”, Last as “ਲਾਸਟ”, Date as “ਡੇਟ”etc. There are a lot of such English words which are very frequently used by Punjabi speaking people even if they are communicating in Punjabi. When a web query is entered by using such words in Gurmukhi script, the available search engines are not capable enough to understand the query and they return irrelevant results. Let’s suppose a Punjabi-English mixed language web query such as:-

ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਅਡਮਿਸ਼ਨ ਦੀ ਲਾਸਟ ਡੇਟ

The above query is written in Gurmukhi script but it is not a Punjabi query. Instead it is a mixed language query because it uses English words admission (ਅਡਮਿਸ਼ਨ), last ਲਾਸਟ) and date (ਡੇਟ) in Gurmukhi script.

Such Punjabi-English mixed language web queries can be processed using NLP to generate a unilingual query. The target language can be English, which will provide more relevant results by searching the information from the global web. This paper presents the methodology to process Punjabi-English mixed language web queries using various modules which are shown in FIGURE 1.

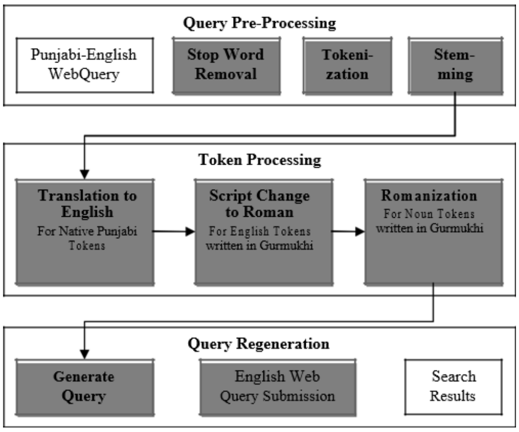

FIGURE 1. Various modules to handle Punjabi-English Mixed Web Query

Query Pre-Processing

The input query Punjabi-English mixed language query in Gurmukhi script is pre-processed to remove any stop words [17]. The stop words are those words that are embedded in the sentence to make it grammatically correct and complete. But when we process a natural language sentence, the removal of these words makes the further processing easier by reducing the number of words to be processed.

After removal of stop words, the query sentence retains only the important words for web search. These words are separated to generate a list. The process of tokenization splits the query sentence into individual words and stores them in an array. These words are called tokens and each token is stored as an element at a separate index. So, the words can be processed individually.

The tokens are stemmed to remove any suffixes. Like other languages, in Punjabi suffixes are used to make plurals and other inflected words. Removal of suffixes provides singular and other stem words. By using suffixes, many variants of the same basic word are produced. After removal of suffixes, the words can be more efficiently processed. It also overcomes the redundant efforts for processing similar words.

Token Processing

A Punjabi-English dictionary was prepared using IndoWordNet. IndoWordNet is a WordNet of Indian languages [18]. The native Punjabi words stored as tokens in the array are translated to equivalent English words. In this step, only native Punjabi words that are available in dictionary get processed. The English words and Punjabi nouns remain unprocessed in the array.

In the second step, the English words that are written in Gurmukhi script are processed. The English words that are frequently used by Punjabi speakers were collected. A Gurmukhi-Roman list was manually prepared containing English words typed in both scripts i.e. Gurmukhi and Roman. Some words are shown in the TABLE 1.

TABLE 1. Some English words typed in Gurmukhi and Roman

English Word in Roman	English Word in Gurmukhi
Admission	ਅਡਮਿਸ਼ਨ
Late	ਲੇਟ
Date	ਡੇਟ
Last	ਲਾਸਟ
Rest	ਰੈਸਟ

The Gurmukhi scripted word is searched in the list and equivalent Roman English word is picked to replace the token.

In the third step Punjabi nouns are processed. At this step, the Gurmukhi nouns are transliterated to Roman script. These are the tokens that remain unprocessed in the first and second step. A rule based algorithm is used to Romanize the Gurmukhi nouns [19].

Query Regeneration

The token processing module generated English tokens (Roman Script) from the input mixed-language tokens (Gurmukhi Script). These English tokens are used to generate a monolingual web query by concatenating all the tokens separating them with a white space. The resultant string is a web query in pure English language that is submitted to the search engine to get the results from the global web. ALGORITHM 1 shows the working of the system.

ALGORITHM 1: Working of the System

```
Let,    Qry(PE): Punjabi-English Mixed Language Query
        Qry(E): Web Query in English
        Dic(PE): Punjabi-English Dictionary
        Lst(GR): Gurmukhi-Roman List
        Lst(T): Token List
        Lst(SW): Stop Words List
        Tkn: Token
        Wsp: White Space
        Tkn(E): Token in English
        Tkn(P): Token in Punjabi
Read Qry(PE)
Search for a Match of Words in Qry(PE) and Lst(SW)
If Matched, Replace that word with Wsp
Tokenize Qry(PE) to generate Lst(T)
Open Lst(T)
While Read(Tkn(P) from Lst(T)) is True
        Stem Tkn(P)
        Update Tkn(P) to Lst(T)
End While
Close Lst(T)
Open Lst(T)
While Read(Tkn(P) from Lst(T)) is True
        Search Tkn(P) in Dic(PE)
        If Matched Then
             Read Tkn(E) from Dic(PE)
             Replace Tkn(P) with Tkn(E) in Lst(T)
        End If
End While
Close Lst(T)
Open Lst(T)
While Read(Tkn(P) from Lst(T)) is True
        Search Tkn(P) in Lst(GR)
        If Matched Then
             Read Tkn(E) from Lst(GR)
             Replace Tkn(P) with Tkn(E) in Lst(T)
        End If
End While
Close Lst(T)
Open Lst(T)
While Read(Tkn from Lst(T)) is True
        If Tkn is Gurmukhi Token Then
             Tkn(E) = English_Transliteration (Tkn)
             Replace Tkn with Tkn(E) in Lst(T)
        End If
End While
Close Lst(T)
Qry(E) = Concatenate_Tokens(Lst(T))
Search Web using Qry(E)
Show Search Results
End
```

TEST AND RESULTS

A list of 407 Punjabi-English mixed language web queries was prepared in Gurmukhi script. The queries contain English words in Gurmukhi script that are frequently used by Punjabi speakers. These queries were processed using the presented methodology and generated English queries were compared with expected queries. The results were promising. TABLE 2 shows some test queries (Punjabi-English mixed language queries) and resultant regenerated queries (English queries) produced by the system.

TABLE 2. Punjabi-English Mixed Language Queries and Regenerated Queries.

Punjabi-English Mixed Language Query	Regenerated English Web Query
ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਅਡਮਿਸ਼ਨ ਦੀ ਲਾਸਟ ਡੇਟ	Punjabi university Patiala admission last date
ਕੋਵਿਡ-19 ਵੈਕਸੀਨ ਦੀਆਂ ਕਿਸਮਾਂ	Covid-19 vaccine types
ਦਿੱਲੀ ਦੇ ਅਨਲੌਕ ਦੀ ਡੇਟ	Delhi unlock date
ਪੰਜਾਬ ਵਿੱਚ ਕੋਵਿਡ ਕੇਸਾਂ ਦੀ ਸਥਿਤੀ	Punjab covid cases status
ਸਿਕਸਥ ਪੇ ਕਮਿਸ਼ਨ ਦੀ ਰਿਪੋਰਟ	Sixth pay commission report
ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੇ ਲੇਟ ਅਡਮਿਸ਼ਨ ਦੇ ਰੂਲ	Punjabi university Patiala late admission rules
ਵਾਟਰ ਦੀਆਂ ਕਿਸਮਾਂ	Water types

Total 364 queries were generated as expected and it gave 89.43% accuracy as shown in FIGURE 2.

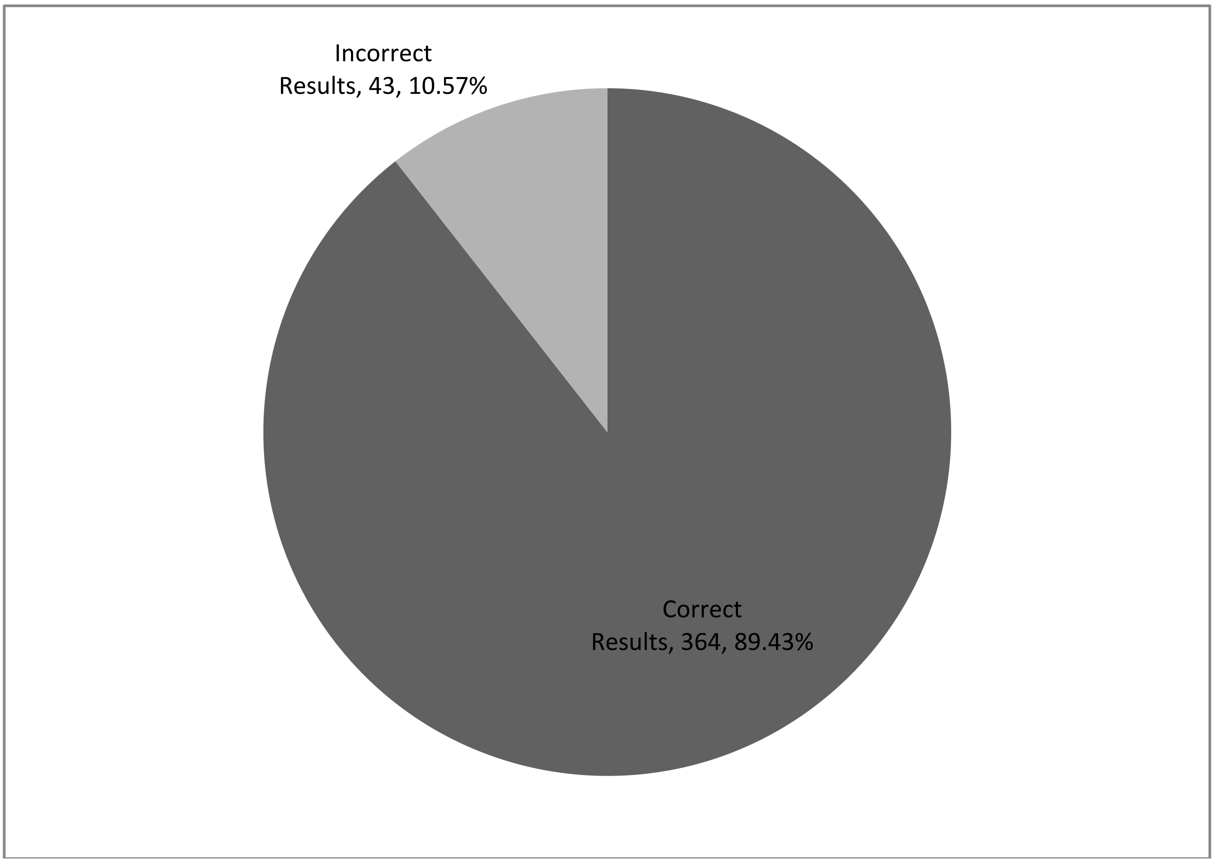

FIGURE 2. Accuracy of Results

The incorrect results were analyzed and it was found that some English words used in query were not available in Gurmukhi-Roman list. By adding more words to this list, accuracy can be further improved.

CONCLUSION

A methodology is presented to process Punjabi-English mixed language web queries. These types of queries are frequently used by Punjabi speaking people in oral as well as written communication. The objective is to generate equivalent English web query. The native Punjabi words are translated to English using a Punjabi-English dictionary, the Gurmukhi scripted English words were replaced with Roman script using a Gurmukhi-Roman list. The remaining words are Punjabi nouns, so they are transliterated to Roman using rule based technique. Total 407 Punjabi-English web queries are used to test the methodology. It gave 89.43% accuracy. The accuracy can be further improved by adding more words to Gurmukhi-Roman list.

REFERENCES

1. Hillier, M.: The role of cultural context in multilingual website usability. Electronic Commerce Research and Applications 2(1), 2-14 (2003). https://doi.org/10.1016/S1567-4223(03)00005-X

2. Miraz Dr., Ali M., Excell P.: Multilingual Website Usability Analysis Based on an International User Survey. In: Proceedings of the Fifth International Conference on Internet Technologies and Applications (ITA 13), pp. 236-244. Creative and Applied Research for the Digital Society (CARDS), Glyndŵr University in Wrexham, North East Wales, UK (2013).

3. M. G. Abbas Malik, "Punjabi Machine Transliteration", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, July 2006, pp.1137–1144.

4. Tariq Rahman , "Soft Power of Punjabi: Language in the domain of Pleasure", Journal of Punjab and Sikh Studies, Vol.24, Nos.1 & 2 (Spring-Fall 2017), pp.73-94.

5. Tej K. Bhatia, "Major Regional Languages", Languages in South Asia, Braj B. Kachru, Yamuna Kachru, S.N. Sridhar (eds.), Cambridge University Press, 2008, pp.121-131.

6. Anne Murphy, "Writing Punjabi across borders", South Asian History and Culture, Vol.9, Issue 1, 2018, pp.68-91, doi:10.1080/19472498.2017.1411049

7. Gyöngyi Rózsa, Anita Komlodi, and Peng Chu: Online Searching in English as a Foreign Language. In: Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), pp. Article 86, 875–880. Association for Computing Machinery, New York, NY, USA (2015). DOI:https://doi.org/10.1145/2740908.2743007

8. Hengyi Fu: Analysis of chinese-english mixed language query reformulation strategies and patterns during web searching. In: Proceedings of the Association for Information Science and Technology, pp. 1–5. American Society for Information Science, USA (2016).

9. Xiaoyi Ma , Mark Y. Liberman, "BITS: A method for bilingual text search over the web (1999)", In Proceedings of the Machine Translation Summit VII, 1999

10. Wei Gao, John Blitzer, Ming Zhou, Kam-Fai Wong, "Exploiting Bilingual Information to Improve Web Search", Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1075–1083, Suntec, Singapore, 2-7 August 2009

11. K. L. Kwok. 2000. Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. In Proceedings of the fifth international workshop on on Information retrieval with Asian languages (IRAL '00). Association for Computing Machinery, New York, NY, USA, 173–179. DOI:https://doi.org/10.1145/355214.355239

12. Jin Zhang, Suyu Lin: Multiple Language Supports in Search Engines. Online Information Review (ISSN: 1468-4527 4527) 31(4), 516-532 (2007). https://doi.org/10.1108/14684520710780458

13. Fotis Lazarinis, Jesus Vilares Ferro, and John Tait: Improving non-English web searching (iNEWS07). ACM SIGIR Forum 41(2), 72–76 (2007). https://doi.org/10.1145/1328964.1328977

14. Lewandowski, D.: Problems with the use of web search engines to find results in foreign languages. Online Information Review 32(5), 668-672 (2008). https://doi.org/10.1108/14684520810914034

15. Wei Gao, John Blitzer, and Ming Zhou: Using English information in non-English web search. In: Proceedings of the 2nd ACM workshop on Improving non english web searching (iNEWS '08), pp. 17–24. Association for Computing Machinery, New York, NY, USA (2008).DOI: https://doi.org/10.1145/1460027.1460031

16. Mequannint Munye and Solomon Atnafu: Amharic-English bilingual web search engine. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES '12), pp. 32–39. Association for Computing Machinery, New York, NY, USA (2012).DOI:https://doi.org/10.1145/2457276.2457284

17. Kaur J, Saini JR: Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle. In: Proceedings of the ACM Symposium on Women in Research 2016 (WIR '16), pp. 32–37. Association for Computing Machinery, New York, NY, USA (2016).

18. Bhattacharyya P.: IndoWordNet. Dash N., Bhattacharyya P., Pawar J. (eds.) The WordNet in Indian Languages, pp. 1–18. Springer, Singapore (2017). https://doi.org/10.1007/978- 981-10-1909-8_1

19. Harjit Singh, Ashish Oberoi: An Efficient Romanization of Gurmukhi Punjabi Proper Nouns for Pattern Matching. International Journal Of Recent Technology And Engineering 8(3), 634-640 (2019). https://doi.org/10.35940/ijrte.B2467.09831