Algorithm for Punjabi Question Answering System

Poonam Gupta  
ME Research Scholar Computer Science & Enng.  
University Institute of Engg and Tech.  
Panjab University, Chandigarh, India

Vishal Gupta  
Assistant Professor, Computer Science &Engg.Dept.  
University Institute of Engg. & Tech.  
Panjab University, Chandigarh, India

Abstract—In this paper an algorithm for Punjabi Question Answering system has been implemented. The proposed system is designed and built in such a way that it increases the accuracy of Question Answering Systems in terms of Recall and Precision and is working for factoid questions and answers text in Punjabi. The system constructs a novel pattern finding and matching system to identify most accurate probable answer out of multiple answers. This algorithm is based on understanding the meaning of the given Punjabi question and expresses them in query logic language. The answers are extracted for the questions of type क्यों(what), जब(when), जहाँ(where), कौन(who) and क्यों(why).

Keywords—Natural Language Processing, Punjabi Question Answering System, Information retrieval.

I. INTRODUCTION

In today’s context, there are three components of Question answering system (QAS) are question classification, information retrieval, and answer extraction. These components play a vital role in QAS. Question classification play chief role in QA system to categorize the question based upon its type. Information retrieval method is to identify success by extracting out applicable answer post by their intelligent question answering system [4]. In comparing with classical information retrieval, where whole documents are considered similar to the information demand, in question answering, specific pieces of information are returned as an answer. The user of a question answering system is interested in a concise, comprehensible, correct and most probable answer, which may refer to a word, sentence, paragraph, image, audio fragment, or an entire document [1]. Question answering system is an important application of the text mining [2]. These systems extract more relevant answers to a particular question posed to them from the large collection of text corpus [5]. On the other hand, various relevant answers can also be extracted classifying the questions into various categories and hence relevancy can be detected easily from the questions and the answers [6]. The Internet today has to face the difficulty of dealing with multi linguism [3]. All the work in Question answering system is done for various other languages but as per knowledge, limited work is done for Punjabi Language.

II. ALGORITHM FOR PUNJABI QUESTION ANSWERING SYSTEM

The proposed algorithm takes Punjabi question and a paragraph text as input from which answers are to be extracted. The algorithm proceeds by segmenting the Punjabi input question into Words. For each word in the question follow following steps:

**Step 1 Query Processing:** - In this step, it classify the questions based upon its categories like बैठें (WHO), क्यों(WHAT), क्यों(WHY), जहाँ(WHERE), जब(WHEN) and the system checks the corresponding rules.

**Step 2** If the input question consists of बैठें (WHO) then go to the procedure for बैठें (WHO).

**ElseIf** the input question consists of क्यों(WHAT) then go to the procedure for क्यों(WHAT).

**ElseIf** the input question consists of जहाँ(WHERE) then go to the procedure for जहाँ(WHERE).

**ElseIf** the input question consists of क्यों(WHY) then go to the procedure for क्यों(WHY).

Else Algorithm cannot find the answer.

**Algorithm Input Example:**

भारत में वार्षिक दास्तांत्रिक में जनसंख्या बढ़ गई। अन्य भारतीय देशों में भी जनसंख्या बढ़ गई। 

**Algorithm Output Example:**

- **WHO:** भारत में वार्षिक दास्तांत्रिक में जनसंख्या बढ़ गई।
- **WHAT:** अन्य भारतीय देशों में भी जनसंख्या बढ़ गई।
- **WHY:** भारत में वार्षिक दास्तांत्रिक में जनसंख्या बढ़ गई।
- **WHERE:** अन्य भारतीय देशों में भी जनसंख्या बढ़ गई।
- **WHEN:** भारत में वार्षिक दास्तांत्रिक में जनसंख्या बढ़ गई।
- **WHICH:** भारत में वार्षिक दास्तांत्रिक में जनसंख्या बढ़ गई।
Questions asked from the above paragraph as below:

1) कवच आॅिंग्रृज़ा बार्त छिड़्य?
2) कवच आॅिंग्रृज़ा चा नएवु वटे एहे निंचे देखिए?
3) कवच आॅिंग्रृज़ा चा धुकु रा जी मी?
4) कवच आॅिंग्रृज़ा अंदटीला चे वटु लघम दिंडु आळे?

Output:-
कवच आॅिंग्रृज़ा बार्त दे अथवा दिंडु मट।
अथ रा नवू 2 अथवा, 1869 श्री, दुि पेक्षुन्टस (अटीबावाजु), नासु दिंडु देें।
अथ रा धुकु रा भेल्स राम वटु लट दां आॅिंग्रृज़ा मी।
अथ अंदटीला चे वटु आॅिंग्रृज़ा दे अथावे वटु लघु लघु मंधम वटु लघम आळे।

A. Procedure for बैंड (WHO)
This procedure is used for finding the named entities from input paragraph. The algorithm for बैंड(Who) has been given below:

Step 1 Extract the substring before बैंड (WHO) from the given question and say it is S1.
Step 2 Extract the substring after बैंड (WHO) upto the end of the question and say it is S2.
Step 3 Initialize Answer = null.
Step 4 If S1 ≠ null then depending upon S1 set the answer
    If S1 contains the name of the person
        // Name is identified by searching current word in the Punjabi Dictionary if it is not in the Punjabi Dictionary and found in the Punjabi name list(it consists of 18000 names) then it is a name//
        { Search the paragraph by matching the strings in S1 and S2 and append the matching details to the end of S1. Say it is S1'. Append S2 after S1' and store it in the Answer.
    }
End of Step 4 loop else
Step 5 If S1 contains the words तुषी (tusī), तुसानूं (tuhānūं), तूं (tūं) then
    { Replace that word as follows:
        तुषी (tusī) to मै (mai)
        तुसानूं (tuhānūं) to मै (mainū)
        तूं (tūं) to मै (main).
    }
    Search the name of person, thing or concept in the paragraph based upon S1.
    // Name can be identified by searching in the Punjabi dictionary if it is not in the dictionary and it is found in the Punjabi name list(it consists of 18000 names) then it is considered as a name// and append it after S1. Say it is S1'. Change the last word of S2 according to the words in S1.
        { If तुषी (tusī) then last word is घे (hō) changes to सां (hāं).
            If तुसानूं (tuhānūं) then last word is same as in substring S2. If तूं (tūं) then last word is घे (hai) changes to सां (hāं) and store it in S2'.
        }
    Append S2' to the end of S1' and store it in Answer.
The procedure has got two rules out of which one of the rule must satisfy to generate a final result. The Rule_1 is if the substring before ਕੀ (WHAT) is null then it stores ਸਾਂ / ਨਸੀ (Yes/No) in Answer. The Rule_2 is if the substring before ਕੀ (WHAT) is not null then the question is of the form “ਤੁਸੀ ਕੀ ਖ੃ਡ ਰਸ੃ ਹੋ?” And the result generated
contain the pattern “भागी द्विजरूप बने होंगे” asī phūṭbāl khē rahē hāṃ.” It can be “भुज्राज आं सो है? tuhāḍā kī nāṃ hai?” And the result generated contain the pattern “भेंज चतुर्द्विध लां दे। mērā carandīp nāṃ hai.” It can be “तूं ली बल लां दे? tūṃ kī kār rahē hai?” And the result generated contain the pattern “भैं किभ कर रहस् hā.”

C. Procedure for क्यों (WHEN)

This procedure is used for finding the time and the date expression for a particular question from the input paragraph. The algorithm for क्यों (WHEN) has been given below:

Step 1 Extract the substring before क्यों (WHEN) from the given question and say it is S1.

Step 2 Extract the substring after क्यों (WHEN) up to the end of the question and say it is S2.

Step 3 Initialize Answer = null

Step 4 If S1 = null and S2 containing the words तुसानूं (tuhānūṃ), तुसी (tusī), तुसाडी (tuhāḍī), में (main) then

Replace the word क्यों (WHEN) with the time and date expression and it can be identified if it is in the following formats:

DD/MM/YY like 02/04/97
DD/MM/YYYY like 02/04/1997
YY/MM/DD like 97/04/02
YYYY/MM/DD like 1997/04/02
MM/DD/YY like 05/26/97
MM/DD/YYYY like 05/26/1997
DD.MM.YY like 02.04.97
DD.MM.YYYY like 02.04.1997
String, Numeric like May 26
Numeric, String like 26 May
Numeric : Numeric like 6:45 and append नूं and then store it in S1’.

Replace the words in S2

ए राज्य (मैन) (mainūṃ)
ए (tus) to भागी (asī)
एदी (tuhāḍī) to भेंजी
में (main) to दूं (tōṃ)
Store it in S2.

Append S2’ to the end of S1’ and store it in Answer.

(end of step 4)

else

Step 5 If S1 ≠ null and is containing the word तुसाडी (tuhāḍī)

Then Replace the word तुसाडी (tuhāḍī) to भेंजी. Store it in S1’.

Replace the word क्यों (WHEN) with the time and date expression and it can be identified if it is in the following formats:

DD/MM/YY like 02/04/97
DD/MM/YYYY like 02/04/1997
YY/MM/DD like 97/04/02
YYYY/MM/DD like 1997/04/02
MM/DD/YY like 05/26/97
MM/DD/YYYY like 05/26/1997
DD.MM.YY like 02.04.97
DD.MM.YYYY like 02.04.1997
DD-MM-YY like 02-04-97
DD-MM-YYYY like 02-04-1997
String, Numeric like May 26
Numeric, String like 26 May

Gupta et al., International Journal of Advanced Research in Computer Science and Software Engineering 3(7), July - 2013, pp. 902-909

© 2013, IJARCSSE All Rights Reserved

Page | 906

This procedure has one rule that must satisfy to generate a final result. The Rule is if the substring before ਕਦੋਂ (WHEN) is null then replace the word ਕਦੋਂ (WHEN) with the time and date expression. The question is of the form “ਕਦੇ ਕਿਮੀ ਹਿਸਚੀ ਦੇ ਨਹਾਨੀ? ਕਦੋ ਤੁਸੀ ਇਸ ਤੋ ਜਾਵਗੀ?” And the result generated contain the pattern “ਕਲ ਅਸੀ ਇਸ ਤੋ ਜਾਵਗੀ।” The question can be of the form “ਕਦੇ ਉਤਲੀ ਉਤਲੀ ਦੇ ਤਿੱਥੋ ਤੋ ਜਾਵਗੀ?” And the result generated contain the pattern “02-04-2012, ਨੂ ਭ੃ ਤਿੱਥੋ ਹੋ ਜਾਵਗੀ।” The question can be of the form “ਕਦੇ ਭਈ ਦੀ ਮਿਦਿਠ ਬਲ ਮਿਤਰਾਲੀ? ਕਦੋ ਮੀਨ ਸੋ ਧੀ ਹੋ ਜਾਵਗੀ।” And the result generated contain the pattern “੫-੧੨-੨੦੧੨ ਮੀ ਮੀ ਬਲ ਮਿਤਰਾਲੀ।” And the question can be of the form “ਕਦੇ ਮੀ ਮਿਦਿਠ ਬਲ ਮਿਤਰਾਲੀ? ਕਦੋ ਮੀਨ ਸੋ ਧੀ ਹੋ ਜਾਵਗੀ।” And the result generated contain the pattern “04/12/2011 ਮੀ ਮੀ ਬਲ ਮਿਤਰਾਲੀ।” And the question can be of the form “ਕਦੇ ਸਾਂ ਬਲ ਸਵਾਲ? ਕਦੋ ਮੀਨ ਸੋ ਧੀ ਹੋ ਜਾਵਗੀ।” And the result generated contain the pattern “04/12/2011 ਮੀ ਮੀ ਬਲ ਮਿਤਰਾਲੀ।”

D. Procedure for ਲਖੀਂ (WHERE)

This procedure is used for finding the location name for a particular question. The Algorithm for ਲਖੀਂ (WHERE) has been given below:

Step 1 Extract the substring before from the given question and say it is S1.
Step 2 Extract the substring after ਲਖੀਂ (WHERE) upto the end of question and say it is S2.
Step 3 Initialize Answer = null.

Step 4 If S1≠ null and S2≠ null and S1 is containing the words ਤੁਝੀ (tusī), ਤੁਸਾਨੂੰ (tuhānū) then

{Replace the words as follows
 ਤੁਝੀ (tusī) to ਮੈ (mainī)
 ਤੁਸਾਨੂੰ (tuhānū) to ਭੂਨੂੰ (mainū) and store it in S1’.

Replace the word ਲਖੀਂ (WHERE) with the matching locations like ਇਨ (In), ਆਟ (At), ਨਹਿਰੇ (Near) from the paragraph or name of the location can also be identified by searching it in the Punjabi dictionary if it is not in the dictionary then it is considered as a location name and store it in S1’.

Change the last word of S2 according to S1
{If ਤੁਝੀ (tusī) then last word is ਸਾ ਦੇ ਹਾ ਸੁਨੀ (sānī).
If ਤੁਸਾਨੂੰ (tuhānū) then last word is same as in substring S2 and store it in S2’.
}
}(end of step 4 loop)

else

Step 5 If S1 is not containing the words ਤੁਝੀ (tusī), ਤੁਸਾਨੂੰ (tuhānū) then

{Keep S1 and S2 as such and store it in S1’ and S2’ respectively.
Just Replace ਵਿਚੇ (WHERE) with the matching locations
like ਬਾਂਚੇ (In), ਮੇਂ (At), ਸੈਂ (Near) from the paragraph or
name of the location can also be identified by searching it in
the Punjabi dictionary if it is not in the dictionary then it is
considered as a location name and store it in S1’.
Append S2’ to the end of S1’ and store it in answer.

This procedure has one rule that must be satisfied to generate a final result. The Rule is if the substring before and after
ਕਕਉ (WHY) is not null then replace the word ਵਿਚੇ (WHERE) with the matching locations like ਬਾਂਚੇ (In), ਮੇਂ (At), ਸੈਂ (Near). The question is of the form “ਤੁਕਿੰਦਰ ਕਕਉ ਲਗਾ ਰਹੇ ਹੋ? cel tuhānūṃ kaidtē mailēgā?” and the result generated contain the pattern “ਦੰਗਾਵਾਲਾ ਕਕਉ ਲਗਾ ਰਹੇ ਹੋ? cdilī dā rahaiṇ vālā hai.”

E. Procedure for ਵਿਚੇ (WHY)
This procedure is used for finding the reason for ਵਿਚੇ (WHERE) type of questions from the input paragraph. The algorithm for ਵਿਚੇ (WHY) has been given below:

Step 1 Extract the substring ਵਿਚੇ (WHY) before from the given question and say it is S1.
Step 2 Extract the substring after ਵਿਚੇ (WHY) upto the end of question and say it is S2.
Step 3 Initialize Answer = null.
Step 4 If S1 ≠ null and S2≠ null then
  { 
    Keep S1 as such.
    Replace the word ਵਿਚੇ (WHY) with some reason by
    matching the strings in S1 and S2 with the paragraph.
    Store it after S1.
    Keep S2 as such.
    Append S2 to the end of S1 and store it in the Answer.
  } (end of step 4 loop)
else

Step 5 If S1 ≠ null , S2≠ null and S1 is containing the words ਤੁਹਾਡੇ (tuhāḍē), ਤੁਸੀ (tusī)
  { 
    Replace the words as follows
    { 
      ਤੁਹਾਡੇ (tuhāḍē) to ਮੇਂ (mērē)
      ਤੁਸੀ (tusī) to ਅਤੀ (asī)
    }
    Replace the word ਵਿਚੇ (WHY) with some reason by
    matching the strings in S1 and S2 with the paragraph. Store
    it in S1.
    Keep S2 as such.
    Append S2 to the end of S1 and store it in the Answer.
  } (end of step 5 loop)

This procedure has one rule that must be satisfied to generate a final result. The Rule is if the substring before and after
ਵਿਚੇ (WHY) is not null then replace the word ਵਿਚੇ (WHY) with some reason by matching the substrings of the question.
The question is of the form “ਚੋਕੇ ਨਮਾਡ ਵਿਚੇ ਧਮਾਉਣਾ ਵਿਚੇ ਲਗਾਉਣਾ ਹੋ? baccē jamāt vaicc khush kauv lag rahē san?” and the
result generated contain the pattern “ਚੋਕੇ ਨਮਾਡ ਵਿਚੇ ਧਮਾਉਣਾ ਵਿਚੇ ਲਗਾਉਣਾ ਹੋ? baccē jamāt vaicc khush kamal
dē jammadain tē lag rahē san.”
III. DATASET

TABLE I
DATASET

<table>
<thead>
<tr>
<th>Total no. of Documents</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Words</td>
<td>19410</td>
</tr>
<tr>
<td>Total Sentences</td>
<td>4800</td>
</tr>
<tr>
<td>Total characters</td>
<td>40</td>
</tr>
<tr>
<td>No. of questions asked</td>
<td>4850</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{\text{No. of correct answers}}{\text{Total no. of questions asked}} \times 100 \)

\[ = \frac{3550}{4850} \times 100 = 73\% \]

TABLE II
ACCURACY TABLE

<table>
<thead>
<tr>
<th>Accuracy of Punjabi Question Answering Algorithm for Document 1</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy of Punjabi Question Answering Algorithm for Document 2</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy of Punjabi Question Answering Algorithm for Document 3</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy of Punjabi Question Answering Algorithm for Document 50</td>
<td>85%</td>
</tr>
</tbody>
</table>

Average Accuracy = \( \frac{\sum \text{Accuracy of Punjabi Question Answering Algorithm for all the Documents}}{50} \times 100 \)

\[ = \frac{3664}{50} = 73\% \]

27% errors are due to these types of questions:

- ਤੁਸੀ ਆਪਣੇ ਪੇਪਰ ਦੀ ਹਿੱਸਾ ਦਿੱਕੇ ਜੀਡੀ? tüşī āpaṇē pēpar dī hīsā dīkē jīdī?
- ਕਿਸ ਦੁਆਰਾ ਤੁਸੀ ਦੀ ਹਿੱਸਾ ਵਿ生活环境 ਜਾਂ ਕਿਸ ਦੇ ਹਿੱਸਾ ਵਿ生活环境 ਜਾਂ? kīs dūārā tüşī dī hīsā vījīndōju jān kīs dē hīsā vījīndōju jān?
- ਪੰਜਾਬੀ ਜਤਾ ਜਿਸ ਦੀ ਸਾਰੀ ਪੰਜਾਬੀ ਜਤਾ ਜਿਸ ਦੀ ਸਾਰੀ? pānjvī bersā kīs dī sārī pānjvī bersā kīs dī sārī?
- ਕੀ ਸੋਨਾ ਕਿਸ ਦੀ ਸੋਨਾ? kī sōnā kīs dī sōnā?
- ਕਿਸ ਦੀ ਸੋਨਾ ਕਿਸ ਦੀ ਸੋਨਾ? kīs dī sōnā kīs dī sōnā?

Since we did n’t have the question types ਕਿਸ (what), ਕਿੱਥੇ (where), ਕਿੰਠਾ (who) and ਕਿਸੁ (why). This system can be helpful for developing other NLP applications. For future scope, since , in the current research work, we have used the concept making a hybrid that works in pattern matching (regular expressions) and new proposed answer finding scoring system, which has yielded for better Recall and Precision value, however for future scope we suggest more work must be done to add more types of questions for question classification and more types of answer patterns must be incorporated with scoring system that balances the probability between different natures of answers.

IV. CONCLUSIONS AND FUTURE WORK

The work illustrated here in context of Punjabi Question Answering System is first of its kind. In this paper we have discussed an algorithm for the implementation of Punjabi Question Answering System. Most of the resources used in this implementation are Punjabi Standard abridged dictionary and Punjabi name list. As no work has been done previously in this direction. The focus of the system has been basically on four kind of questions of type ਕੀ (what), ਕਦੋਂ (when), ਕਿੱਥੇ (where), ਕਿੰਠਾ (who) and ਕਿਸੁ (why). This system can be helpful for developing other NLP applications. For future scope, since , in the current research work, we have used the concept making a hybrid that works in pattern matching (regular expressions) and new proposed answer finding scoring system, which has yielded for better Recall and Precision value, however for future scope we suggest more work must be done to add more types of questions for question classification and more types of answer patterns must be incorporated with scoring system that balances the probability between different natures of answers.
ACKNOWLEDGMENT
Many thanks to Mr. Vishal Gupta, Assistant Professor in UIET, Panjab University Chandigarh for doing this work.

REFERENCES