Gravity based Punjabi Question Answering System

Gursharan Singh Dhanjal
M.Tech (IT), BBSB
Engineering College,
Fatehgarh Sahib, Punjab, India

Sukhwinder Sharma
Assistant Professor,
BBSB Engineering College,
Fatehgarh Sahib,
Punjab,INDIA

Paramjot Kaur Sarao
M.Tech (IT), BBSB
Engineering College,
Fatehgarh Sahib,
Punjab,INDIA

ABSTRACT
Question and answers systems are foundation for building intelligent information retrieval systems but, current server of these show that limited effort has been in building Punjabi Question Answer systems. This paper is an attempt to overcome this limitation. The research work is based on a concept taken from physics: Point of Gravity. In this approach the question and answer text are processed to extract numerical features so as to determine Point of Gravity. Matching Gravity Score values are computed for finding answer against a question query. Series of random evaluation sample sets show a high degree of overall system accuracy (above 91%) for each question type in terms of precision and recall. Individually the evaluation of each question type also shows not less than 91% accuracy in terms of precision.

General Terms
Information retrieval, natural language processing, question answering, gurmukhi script

Keywords
Information retrieval, natural language processing, question answering, gurmukhi script

1. INTRODUCTION
It has been observed that in recent past, the focus of these researchers has been limited to few types of questions: ਕੌਣ (Who), ਕੀ (What), ਕਦੋਂ (When), ਕਿੱਥੇ (Where) & ਕਿਉਂ (Why). Either of these approaches use regular expressions based searching methods or they are based on statistical methods. However some researchers have used a hybrid or a combination of these two approaches. All these approaches are essentially scoring systems, based on which the selection of best possible answers is done.

However, when these Punjabi Question Answering (QA) systems are evaluated in current context in terms of technology stack, it can be observed that most of the previous works / spade work done in this area of Punjabi QA systems is elementary. They are not using cloud based technologies, open-source approaches so that can be used easily by any user or extended for further research. Nor they are using recent advancements achieved in the domain of Natural Language Processing.

Therefore, there is an ample scope not just to increase the coverage of existing Punjabi QA systems, but also to upgrade the technology stack and improve their accuracy.

2. RELATED WORK
Kumar et al. [1], developed a Hindi QA system whose goal was to help elementary and high school students to find correct answers for questions from subjects. A no word-net was available in Hindi so they constructed a lexical database of synonyms. Questions are classified through case based rules classifies the question, change to proper query and submitted to retrieval engine.

Sekine et al. [2] implemented a cross-lingual question-answering (CLQA) system for Hindi and English. Here, input is an English question and candidate answers are searched from Hindi newspapers, which are further translated into English with the context of each answer. Top 20 articles used were extracted using an English Hindi bilingual dictionary.

A Hindi QA system, "PRASHNOTTAR" is implemented by Sahu et al. [3]. The parser performs morphological analysis, chunking and POS tagging. Four question classed were used: "when", "where", "what time" and "how many" including 15 questions of each class.

Stalin et al. (2012) [4] implemented a web based application for finding answers of Hindi language questions from Hindi text. If Hindi Text does not contain the answer, then it is searched through Google search engine. The author proposed an architecture that used question words and pattern of the question to extract all candidate answers from the retrieved documents.

A Question Classification System in Bengali language (and eastern Indic Language) is implemented by Banerjee et al. [5]. Author(s) extended their own work to propose two-layer taxonomy with 69 fine-grained & 9 course-grained classes.

Pakray [6]: This research followed a SQL based approach for finding answers for the respective question. The system has a Dialogue Manager which helps in maintaining the missing information if any while executing 'Query' formed from the question. Technically, the system tags the dialogue questions in a semantic manner based on shallow parser for a specific domain.

Reddy et al. [7]: made a QA system that works in the domain of railways. The questions and answers are based on Telegu language keywords. System analyzes the query (Question) in terms of tokens from which it can identify, what is being asked for. Possible answers are displayed to the user using SQL queries generated in the system.

Gupta et al. [8] used a regular expression and ranking algorithm to implement a Punjabi QA. The only existing Punjabi QA system included the following question types: ਕੀ (what), ਕਦੋਂ (when), ਕਿਉਂ (why), ਕੌਣ (who) and ਕਿੱਥੇ (where). The processing of question answer text is done with using stop word removers, stemmers [9] [10], regular expressions and an algorithm for Question & Answer patterns [11]. Question type and possible answers are extracted using a pattern matching approach. Possible answers are ranked using a scoring system, which provides the most appropriate answer of the given question with highest score.

A tabular comparison of the all the systems discussed above is presented in Table 1.

Table 1: Comparison of Indian Language QA Systems

| Paper			Question Frequency	Evaluation	Demerits
Praveen Kumar et al. [1]			150	Accuracy (59%)	Unsuitable for bigger systems due to limited knowledge data.
Shriya Sahu [3]			60	Accuracy (68%)	No use of semantics and dynamic data set.
Rami Reddy et al. [7]			95	Precision (96.34%)	Low coverage of the railway domain.
				Dialogue Success Rate (83.96)	
Somnath Banerjee et al. [5]			1100	Accuracy (87.63%)	Classification is confined to specific language (Bengali).
Shalini Stalin et al. [4]			Sets of 20	Results are inconclusive.	Results are inconclusive.
Partha Pakray [6]	Without Dialogue Management	Bengali	70	Precision (87.50%)	- Small knowledge base due to low coverage of - Faulty tokenization due to language rules.
				Recall (80%)	
		Telegu	132	Precision (97.63%)	
				Recall (93.93%)	
	With Dialogue Management	Bengali	58	Precision (83.67%)	
				Dialogue Success Rate (72.91%)	
		Telegu	62	Precision (96.49%)	
				Dialogue Success Rate (89.06%)	
S. Sekine et al. [2]			56	MRR (25%)	Machine Translation is error prone.
Poonam Gupta et al. [8]			200	Average Precision (85.66%)	- Specific domain. - Types of questions covered are less in number
				Average Recall (65.28%)	
				MRR (0.43%)	

3. SCOPE OF WORK

After conducting a systematic review [12] in context of Punjabi QA systems and based on the identified gaps, the scope of work can defined as follows:

- To build a well formed and structured dataset of Punjabi Comprehensions and Punjabi Question sets.

- To build a Gravity [13] based Punjabi Question Answering System that can cover maximum Question Types, viz. ਕਿਵੇਂ (How), ਕਿੰਨਾ / ਕਿੰਨੀ / ਕਿੰਨੇ / ਕਿੰਨੀਆਂ (How much / How Many), ਕੀ (What), ਕਦੋਂ (When), ਕਿਉਂ (Why), ਕੌਣ / ਕਿਸ / ਕਿਸੇ (Who/Whom/Whose), ਕਿੱਥੇ / ਕਿੱਥੋਂ (Where), ਕਿਹੜੇ / ਕਿਹੜਾ / ਕਿਹੜੀ (Which), ਕਿਹੋ-ਜਿਹਾ / ਕਿਹੋ-ਜਿਹੇ (What kind / What type).

- Evaluate the accuracy of the system using random sampling techniques.

4. METHODOLOGY ADOPTED

This system's working is based on the numerical features extracted from given Question and respective Comprehension. It constitutes the extraction of features such Named Entities, Epistemic Score (based Punjabi language rules), Lexical Density [14], Readability Index [15], Point of Gravity [13] and Matching Gravity Score (Section 4.1) of Question & Sentence text. Fig 1 shows the overall working process of the system and its implementation details are described in Section 4.24.1.
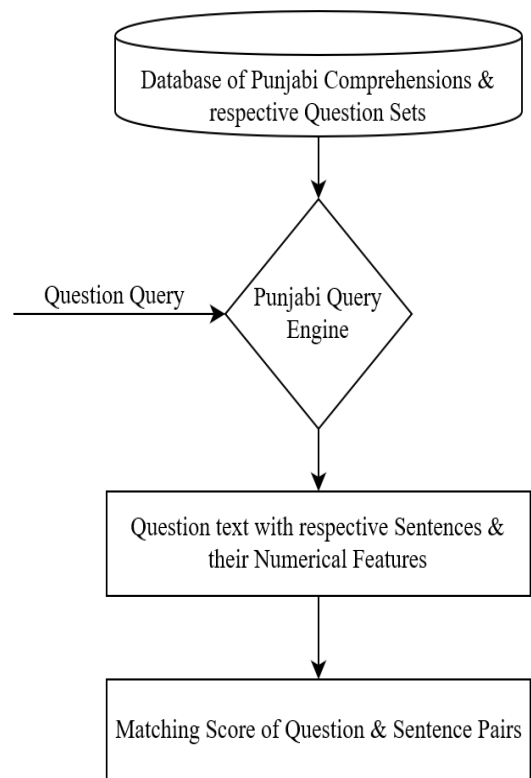
Fig 1: Block Diagram of Punjabi QA System

4.1 Pseudo-logic of Punjabi QA System

Input: Punjabi Question Text

Output: *pair_array* sorted with highest *matching_gravity_score*

Processing:

- foreach *question_query*

  - fetch the respective *comprehension*

  - find sentence boundaries from *comprehension* and build a *sentence_ array*

  - foreach sentence in *sentence_array*

    - apply *filter_rules* to the *sentence_array*

    - build *question_sentence* pairs

    - extract numerical features of each pair

    - calculate *point_of_gravity*

    - *calculate matching_gravity_score* for each pair

    - push pairs into a *pair_array* with their numerical features

  - end foreach

    - sort *pair_array* with highest *matching_gravity_score*

- end foreach

4.2 Implementation Steps

In first phase a database has been designed for storing Comprehensions and their respective Questions. The data set has been collected from Punjabi Text books of 9th and 10th Standard, Previous implementation of Punjabi QA [8], and Articles from "Ajit" newspaper. This set includes nine types of questions, viz. ਕਿਵੇਂ (How), ਕਿੰਨਾ / ਕਿੰਨੀ / ਕਿੰਨੇ / ਕਿੰਨੀਆਂ (How much / How Many), ਕੀ (What), ਕਦੋਂ (When), ਕਿਉਂ (Why), ਕੌਣ / ਕਿਸ / ਕਿਸੇ (Who/Whom/Whose), ਕਿੱਥੇ / ਕਿੱਥੋਂ (Where), ਕਿਹੜੇ / ਕਿਹੜਾ / ਕਿਹੜੀ (Which), ਕਿਹੋ-ਜਿਹਾ / ਕਿਹੋ-ਜਿਹੇ (What kind / What type). The number of questions in each type of question in the data set is: 56, 28, 95, 22, 33, 91, 27, 64 and 22 respectively. The data set is then tagged by Punjabi POS tagger [16] to form a semantic Punjabi dictionary in the database.

When the system is given input as a Question query, it is sent to the Punjabi Query Engine, which processes the Question and its respective comprehension. The comprehension is parsed and relevant sentences are filtered using the Punjabi language rules according to the type of question. Using the semantic Punjabi dictionary build earlier, question and sentence text are analyzed for their numerical features. Following are the numerical features which extracted:

- $X1 = NE\ Frequency$ (1)

- $Y1 = Unique\ NE\ types$ (2)

- $X2 = Ratio\ of\ Intersection$ (3)

- $Y2 = Empirical\ Score$ (4)

- $X3 = Trigram\ score$ (5)

- $Y3 = Difference\ Score$ (6)

- $W1 = W_{lex}/W$ (7)

- $W2 = 0.4\ (W + 100(Compound\ Words/W))$ (8)

- $W3 = W_{lex}$ (9)

Where W is the number of words in the sentence and W_{lex} is the number of lexical words in the sentence.

- $XPG = \frac{(X1 \times W1) + (X2 \times W2) + (X3 \times W3)}{W1 + W2 + W3}$ (10)

- $YPG = \frac{(Y1 \times W1) + (Y2 \times W2) + (Y3 \times W3)}{W1 + W2 + W3}$ (11)

- $Distance = \sqrt{(XPG(Q) - XPG(A))^2 + (YPG(Q) - YPG(A))^2}$ (12)

- $Matching\ Gravity\ Score = \frac{1}{1 + Distance}$ (13)

Where XCG and YCG are using the concept of Point of Gravity [13] to calculate respective coordinates. $XPG(Q)$, $YPG(Q)$ are the values of the given question and $XPG(A)$, $YPG(A)$ are the values of the sentence to matched against it.

The Final list of sentences (probable answers) is ranked according to the *Match Gravity Score* calculated by the system. Sentence having the highest score is the most probable answer for the given question.

5. RESULTS

The performance of the Punjabi QA has been evaluated using random sampling method. Random samples of each question types are in increasing order of 10 were picked and were queried. This section discusses and shows graphically the outcomes from evaluation process.

5.1 Graphs

The "How" graph (Fig 2) shows that with an increase in number of evaluation samples, the precision remains in range bounds (0.93, 1). This range bound seems to be tight and reflects that the standard deviation is low (0.02). From this it can be interpreted that system behavior is consistent for "How" types of questions. The mean precision is 0.958 and its curve is more or less plateau with a slope close to zero, i.e. a nearly horizontal curve with downward trend. This reflects that the precision slightly decreases when number of samples increases after 50. However, when the recall curve is observed, a linearly increasing trend can be seen. The rate of change or the slope of the recall curve is positive. From this it can be interpreted that as the size of the evaluation dataset increases the recall also increases linearly. The recall shows the number of correct answers that have been retrieved from the full evaluation set. The small value of the standard deviation shows that results remain consistent and there is not much difference in worst case and best case of accuracy evaluation.
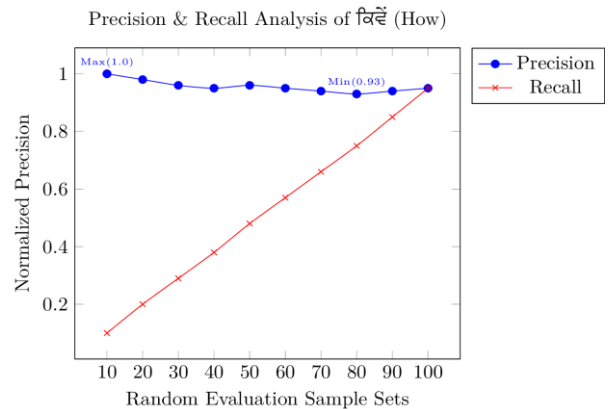

Fig 2: Precision & Recall of question type ਕਿਵੇਂ (How)

In case of "What", it can be observed that the precision never falls below 0.9 and go up 0.98. When the value is 0.98 (Fig 3), both the recall and precision meet because number of samples in both cases become 100. The slope of the precision graph is positive with curve's initial upward movement and then it becomes zero when the curve nearly becomes flat, giving a mean precision of 0.9365. The consistency of correct answers is a bit lower here due relatively higher standard deviation (0.03).
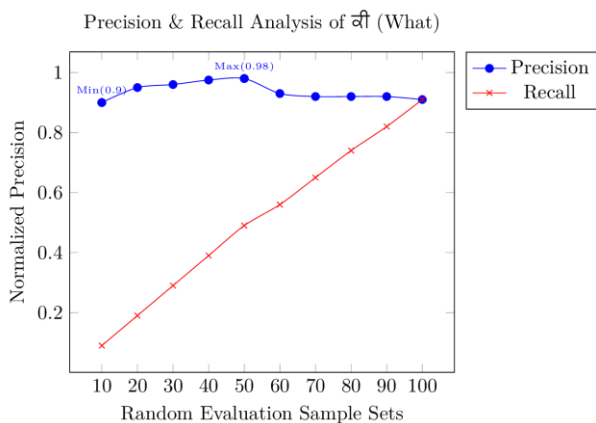

Fig 3: Precision & Recall of question type ਕੀ (What)

In case of "Why" graph (Fig 4), the maximum precision achieved is 0.99 and the minimum is 0.93. There are two cases where the minimum value 0.93 occurs. From this, it can be interpreted that in worst case the value remain close to 0.93 and in best case of sample evaluation the value is 0.99. The precision remains high even with increase in size of the sample. As the sample size increases, the recall also increases. This is because all the samples being evaluated are producing correct results. The Punjabi Question & answer system is able to find 94% correct answers from "Why" type of questions. The standard deviation in precision is also low (0.02), which means that results do not deviate much from the average score. The recall graph, which checks the accuracy based on the complete dataset also show encouraging positive results. There is a point when sample size becomes 100, at this point there is no difference in the recall and precision value.
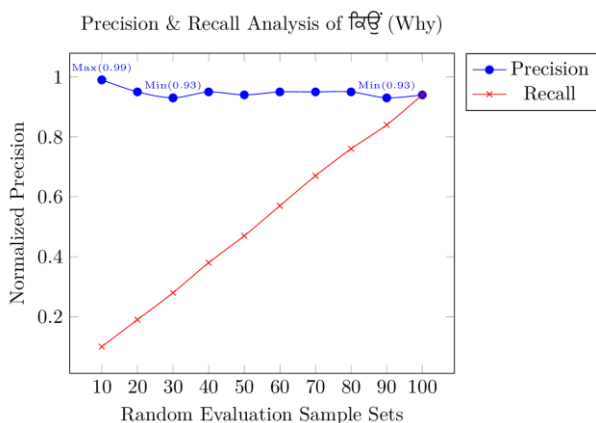

Fig 4: Precision & Recall of question type ਕਿਉਂ (Why)

The Punjabi QA system's results in case of "When" (Fig 5) are highly consistent which is evident from the value of standard deviation (0.01) of precision. This also reflects that rate of change as the number sample size changes, is also low. The recall graph shows linearly increasing trend, which is in

congruence with the precision values. There are two cases when the precision is lowest (0.9). It occurs when the sample size is 10 and 30, other-wise precision remains higher in all other cases of evaluation, which give the mean precision (0.9191).
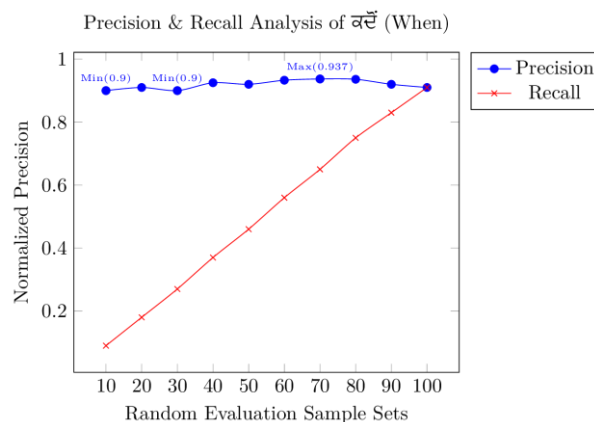

Fig 5: Precision & Recall of question type ਕਦੋਂ (When)

The "Which" graph (Fig 6) shows the curves having behavior similar to the "What" graph. In both cases, the precision is as low as 0.9 in the beginning, then due to upward movement, there is a slight increase in precision. But later on, the precision curve is downwards which is us a mean precision value, 0.9235. The curve also shows a consistent behavior having a low standard deviation (0.01). Also the recall curves have a positive slope which increases proportionally as the samples increase, in both cases.
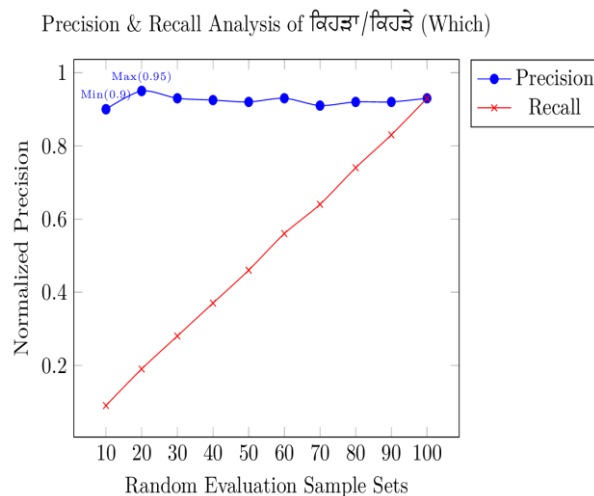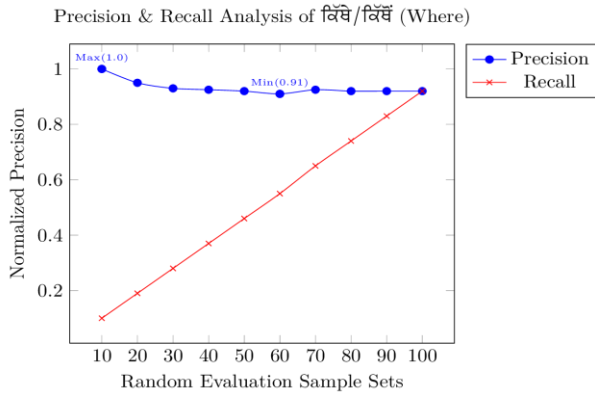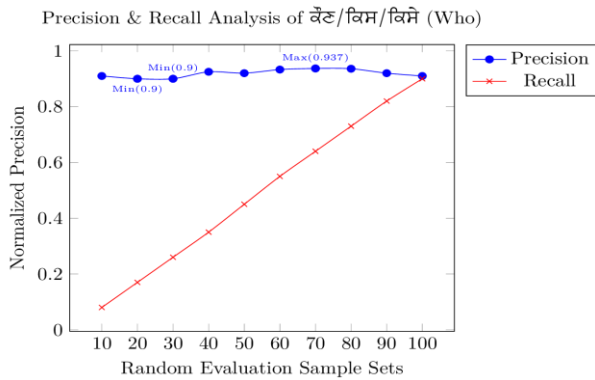

Fig 6: Precision & Recall of question type ਕਿਹੜਾ/ਕਿਹੜੇ (Which)

From the "Where" line graph (Fig 7), it can be observed that initially the precision is 100%, when the sample size is 10. But as the sample size increases, the precision values drop as low as 0.91. The minimum of 0.91 values is also fairly good score in terms of accuracy. In this case, the mean precision provided by the system is 0.932. The standard deviation is bit higher (0.03) than the other cases, "When", "Who", "Where" and "What" and the recall values also clearly follow a positive slope. The recall and precision curves meet at point when sample size is becomes 100, showing that both statistics are in concordance.

Precision & Recall Analysis of ਕਿੱਥੇ/ਕਿੱਥੋਂ (Where)

Fig 7: Precision & Recall of question type ਕਿੱਥੇ/ਕਿੱਥੋਂ (Where)

The "Who" graph (Fig 8) shows relatively lower precision values pointing at a value 0.90 when the sample is size is relatively low and a slightly higher precision value (0.937) when sample size is increased. The standard deviation is also low in this case (0.01) and the mean precision is 0.9191. These value ranges are lowest amongst all the seven types of Punjabi questions. The recall values also maintain the lowest level i.e. 0.9.

Precision & Recall Analysis of ਕੌਣ/ਕਿਸ/ਕਿਸੇ (Who)

Fig 8: Precision & Recall of question type ਕੌਣ/ਕਿਸ/ਕਿਸੇ (Who)

The line graph (Fig 9) of "What Kind / What Type" question type has a similar behavior to the "When" graph (Fig 5), but with a difference that this one has zero Standard Deviation (0.0). This means that this case has a little bit more consistent results with a mean precision value of 0.96. The precision and recall have the ranges similar to "When" graph. This shows that the proposed algorithm (Section 4.1) is producing similar levels of accuracy for both types of questions.

Precision & Recall Analysis of ਕਿਹੋ-ਜਿਹਾ/ਕਿਹੋ-ਜਿਹੇ (What Kind / What Type)
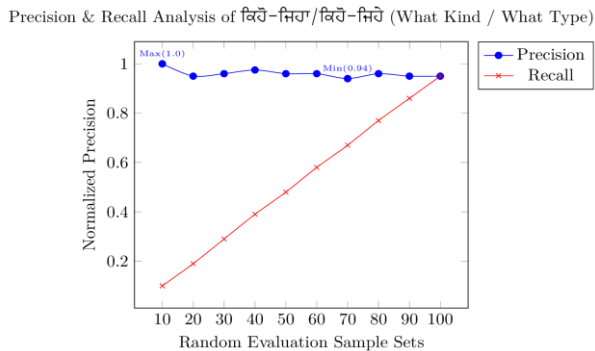
Fig 9: Precision & Recall of question type ਕਿਹੋ-ਜਿਹਾ/ਕਿਹੋ-ਜਿਹੇ (What Kind / What Type)

The graph (Fig 10) of "How much/How many" show that, its precision remains in bracket of 0.94 and 1. From this range, it can be inferred that the Punjabi QA system is producing high quality results (mean precision = 0.9625) and the system never behaves in an erractic manner because of the low standard deviation value (0.02). The recall curve is also shows consonance with respect to the precision values.

Precision & Recall Analysis of ਕਿੰਨਾ/ਕਿੰਨੇ/ਕਿੰਨੀਆਂ (How much / How many)
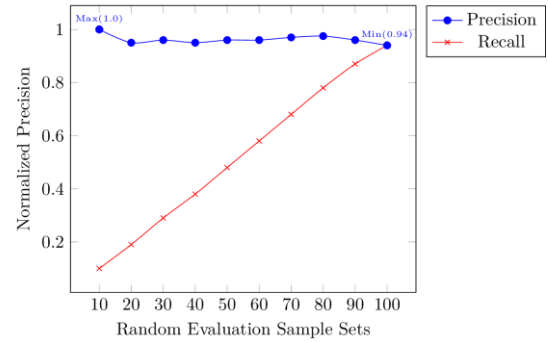
Fig 10: Precision & Recall of question type ਕਿੰਨਾ/ਕਿੰਨੇ/ਕਿੰਨੀਆਂ (How much / How many)

6. DISCUSSION

The key finding of this research work is that Punjabi QA systems are still in an evolving stage. Some, initial spade work has been done by few researchers and this work is next level of iteration for building a Punjabi language based Question Answering system. The work presented here is a novel approach in which uses the Point of Gravity concept to build the Gravity Based Punjabi QA System. The experimental results show a high level of precision for each type of question. The system was evaluated on total of 400 unique questions. It was found that in Who, When types, the mean precision remains above 91%, but in case of What, Why, What, How Many types the mean precision is even higher that 93%. The system is showing high level of consistency as inferred from the low values of standard deviation for each type of question. The previous Punjabi QA system [8], also has high precision and recall values. But this work, covers more types of questions (ਕਿਵੇਂ (How), ਕਿੰਨਾ / ਕਿੰਨੀ / ਕਿੰਨੇ / ਕਿੰਨੀਆਂ (How much / How Many), ਕਿਹੋ-ਜਿਹਾ / ਕਿਹੋ-ਜਿਹੇ (What kind / What type)) and the evaluation of these also yielded high precision and recall values. It was found that the usage of Punjabi Stemmer does not help in increasing the accuracy of the Gravity based Punjabi QA system. The reason for this is the fact that this system relies on lexical features of words, and most of the times their sense changes according to context during the stemming process. Similarly, the usage of Punjabi stop words dictionary does not help in finding or retrieving better answers from the comprehension.

7. CONCLUSION

Following conclusions and inferences are drawn from the performance results of the implemented system.

- The overall accuracy of the Gravity based Punjabi QA System developed is above the average of 91%.

- The system performs well for questions having a single answer.

- The system accomplishes good level of accuracy with all 7 types of Punjabi questions types.

- The retrieval time for finding answers is low.

8. FUTURE SCOPE

In following way this current work can be extended :

- Additional type of Punjabi Questions may be covered like

- Evaluation of the system's performance for next version should also include retrieval time analysis.

9. REFERENCES

[1] P. Kumar, S. Kashyap, A. Mittal, and S. Gupta, "A Hindi Question Answering system for E-learning documents," in *Third International Conference on Intelligent Sensing and Information Processing*, 2005, pp. 80-85.

[2] Satoshi Sekine and Ralph Grishman, "Hindi-English Cross-Lingual Question-Answering System," *ACM Transactions on Asian Language Information Processing*, vol. 2, no. 3, pp. 181-192, September 2003.

[3] Shriya Sahu, Nandkishor Vasnik, and Devshri Roy, "Prashnottar: A Hindi Question Answering System," *International Journal of Computer Science and Information Technology (IJCSIT),* vol. 4, no. 2, pp. 149-158, 2012.

[4] Shalini Stalin, Rajeev Pandey, and Raju Barskar , "Web Based Application for Hindi Question Answering System," *International Journal of Electronics and Computer Science Engineering*, vol. 2, no. 1, pp. 72-78, 2012.

[5] Somnath Banerjee and Sivaji Bandyopadhyay, "Ensemble Approach for Fine-Grained Question Classification in Bengali," in *27th Pacific Asia Conference on Language, Information, and Computation*, 2013, pp. 75-84.

[6] Dr. Partha Pakray, "Multilingual Restricted Domain QA System with Dialogue Management," *Master's Thesis Report, Jadavpur University, Kolkata*, 2007.

[7] Rami Reddy Nandi Reddy and Sivaji Bandyopadhyay, "Dialogue based Question Answering System in Telugu ," in *EACL 2006 Workshop on Multilingual Question Answering - MLQA06*, 2006, pp. 53-60.

[8] Poonam Gupta and Vishal Gupta, "Hybrid Approach for Punjabi Question Answering System," *Advances in Intelligent Systems and Computing*, vol. 264, pp. 133-149, 2014.

[9] Mandeep Singh Gill, Gurpreet Singh Lehal, and Shiv Sharma Joshi, "Part-of-Speech Tagging for Grammar Checking of Punjabi," *The Linguistics Journal*, vol. 4, no. 1, pp. 6-22, May 2009.

[10] Vishal Gupta and Gurpreet Singh Lehal, "Automatic Keywords Extraction for Punjabi Language," *International Journal of Computer Science Issues*, vol. 8, no. 5, September 2011.

[11] Vishal Gupta and Gurpreet Singh Lehal, "Automatic Text Summarization System for Punjabi Language," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 3, August 2013.

[12] Gursharan Singh Dhanjal and Sukhwinder Sharma, "Advancements in Question Answering Systems Towards Indic Languages," *International Journal of Research in Computer Science*, vol. 5, no. 1, pp. 15-26, 2015.

[13] The University of Winnipeg. (1997) Center of Gravity. [Online]. http://theory.uwinnipeg.ca/physics/rot/node4.html

[14] Wikipedia contributors. Lexical density. [Online]. https://en.wikipedia.org/wiki/Lexical_density

[15] Wikipedia contributors. Gunning fog index. [Online]. https://en.wikipedia.org/wiki/Gunning_fog_index

[16] Gurpreet Singh Lehal, Vishal Goyal, and Umrinderpal Singh. Punjabi POS Tagger. [Online]. http://punjabipos.learnpunjabi.org/default.aspx

[17] Vishal Gupta and Gurpreet Singh Lehal, "Named Entity Recognition for Punjabi Language Text Summarization," *International Journal of Computer Applications*, vol. 33, no. 3, pp. 28-32, November 2011.

[18] Wikipedia contributors. Lexical item. [Online]. https://en.wikipedia.org/wiki/Lexical_item