Enhancing Comparator Extraction for Decision Making Process

Shrutika Narayane
M. E. Student
Department of Information Technology
MIT College of Engineering, Kothrud
Pune, India
shrutika1312@yahoo.com

Sudipta Giri
Professor
Department of Information Technology
MIT College of Engineering, Kothrud
Pune, India
sudipta.giri@mitcoe.edu.in

***Abstract* — Comparisons of different things plays crucial role while purchasing in human decision making process. Even though, decision making is common in our day-today life but requires proper sufficient knowledge and skills to know what should get compare and what will be substitutes so as to make good decision. To make decision process easy, in this paper we present a novel way to identify comparative questions and its comparators from user's questions posted online. Also the issues of ambiguous entities raised in comparative questions are addressed. The experimental results on developed system for comparative questions with ambiguous and unambiguous entities are present.**

Keywords- Comparative questions, Entity mining, Bootstrapping Algorithm, Sequential patterns,Inductive Extraction Pattern.

I. INTRODUCTION

All Comparisons of different things are one of the most convincing ways of decision-making which raises substitute options on a daily basis however it requires skill and high knowledge expertise. In case of purchasing things, it becomes difficult for a person with insufficient or poor knowledge to make a better decision to finalize best product according to his need and also making comparison of alternative options. Comparative question and its comparators are two main components of decision making process.

Comparative questions : A question intended to compare two or more entities which are explicitly mentioned in the question archived by users.

***Comparator* :** Target entities in a comparative question which are intended to be compared are comparative entities also called as comparators.

In general, it's conflicting to decide whether two entities are comparable to each other or not since people compare mangos and oranges for various reasons. For example, "Audi" and "BMW" might be comparable products as "car manufacturers" or as "targeted market products". Hence things become wired and complicated when an entity has several parameters and functionalities.

In the following example question Q1 & Q2 are not comparative questions whereas Q3 is comparative question in which

"Mercedes" and "Audi" are comparators.

Q1. "Which one is better?"

Q2. "Is Audi the best car?"

Q3. "Which car is better car Mercedes or Audi?"

The outcomes of such comparative questions will be very useful in helping user's option exploration i.e. recommending Various alternatives choices by suggesting comparable entities on the basis of other previous online user's requests and are just like recommender system which will be very useful for user's choice exploration user's requests.as well as ranging these comparators is beneficial in making decisions. Our attempt is Specially addressing the problem of newer user to finding out good comparators to support his comparison activity, a bootstrapping algorithm is used to identify comparative question and its comparator so as to recommend identified comparators stored in database to new user's those who are unaware about substitute entity available in market to their choice. Experiments are performed on developed system for two and multi entity comparative questions.

The rest of paper is structured as follows. In Section II a short literature survey is given, Section III gives a brief review of implementation of design for extraction and evaluation of inductive extraction pattern while Section IV gives experimental results along with screen shots, conclusion in Section V.

II. LITERATURE SURVEY

Initially most relevant work is done on mining comparative sentences and their relations by Jindal and Liu [1], [15] that is Supervised learning which tends to the machine learning task of containing a function from labeled training sets of data. The training data consist of a set of training examples and uses the class sequential rules (CSR) and label sequential rules (LSR) to identify comparative sentences and extract comparative relations.

CSR is a classification rule which maps a sequence pattern $S\ (s_1, s_2 \ldots s_n)$ (a class *C*. *C* is either comparative or noncompetitive).

LSR maps an input sequence pattern $S\ (s_1, s_2 \ldots s_i \ldots s_n)$ to a labeled sequence $S\ (s_1, s_2 \ldots l_i \ldots s_n)$ by replacing token s_i in the input sequence with a designated label (l_i) and this token is referred as the anchor

But J & L's method have some drawbacks like limited domains and require large amount of keywords indicating comparative sentences. Firstly they manually created a set of 83 keywords like exceed, beat, outperform and better that are indicators of comparative questions. Evaluation of an entity or event is directly comparing it with a similar entity or event [16]. Identification of comparative sentences from texts and to mine comparative relations from its identified comparative sentences, it can achieve high precision but gives low recall.

However, supervised training for exact entity and relation extraction is expensive, requiring a substantial number of labeled training sets for each type of entity and relation to be extracted. Because of this, researchers have

explored semi-supervised learning methods that use small number of labeled examples of the predicate to be extracted, along with a large volume of unlabeled text [16]. Whereas bootstrapping method is very significant one in previous information mining research [7], [9] also referred as weakly supervised bootstrapping technique, significant to extract comparable entities with highly precise manner [2] i.e. with high recall as well as high precision preferred. More details can be found in our previous work presented in [5].

III. IMPLEMENTATION DETAILS

A. Design

Resolving the conflict for extracting comparative questions and its comparator a weakly supervised bootstrapping method is used which will enrich in recall as well in precision [2].

I. Indicative Extraction Patterns Mining

A sequential pattern used for comparative question identification and its comparator extraction is Indicative Extraction Pattern (IEP). There are mainly following three methods used for information extraction as [7], [8], [9] given below,

1. Rule based extraction :

One approach of IE is to automatically learn pattern-based extraction rules for identifying each type of entity or relation. For example, system which was developed by Rapier [10]. Patterns are expressed in an enhanced regular-expression language; and a bottom-up relational rule learner is used to induce rules from a corpus of labeled training examples. Inductive Logic Programming (ILP) [11] has also been used to learn more number of logical rules for identifying phrases to be extracted from a document [12], [13].

2. Pattern based extraction :

Pattern based approaches build on annotated text fragments (the patterns), where words/phrases are labeled with linguistic information, e.g. word lemma, POS-tag or related syntactic information. Patterns which matched against linguistically annotated data to find relationships [14].

3. Supervised learning :

Supervised learning is the machine learning task of inferring a function from labeled training data. And this trained data consist of training examples set. In case of supervised learning, each example is a pair which consists of an input object (typically a vector) and a desired output value (also called the supervisory signal).

A question may matches single or multiple IEP'S and accordingly get classified as comparative a comparative question matches an IEP and the token sequences corresponding to the comparator slots in the IEP are extracted as comparators.

In case of multiple IEPs matching, the longest IEP is used. Therefore we create a set of IEPs automatically; instead of manual inductive keyword creation which is basic idea of weakly supervised method iteratively works as shown in Fig. 1. The two key steps in this algorithm are pattern generation and pattern evaluation.

II. Pattern Generation

Based on two key assumptions mentioned below, bootstrapping algorithm designed as shown in Fig. 1. [3], [16]. If a sequential pattern can be used to extract many reliable comparator pairs, it's most probably to be an IEP.

If a comparator pair can be extracted by an IEP, the pair is reliable. Three types of patterns are generated by using bootstrapping method as given below:

1. Lexical patterns :

Lexical patterns - These patterns indicate sequential patterns consisting of only words and symbols ($C, #start, and #end).

2. Generalized patterns :

A lexical pattern is too specific for matching. So lexical patterns are generalized by replacing one or more words their POS tags.

3. Supervised learning :

Specialized patterns - Pattern specialization is done by adding POS tags to all comparator slots. For example, from the lexical pattern *'<$ or $C>'* and the question 'Paris or London?', *'<$C=NN or $C=NN?>'* will become specialized pattern.

III. Pattern evaluation:

As Bootstrapping procedure gives very few reliable comparator pairs in its early stage. Hence for discovering more reliable pair's, pattern evolution operation is performed.

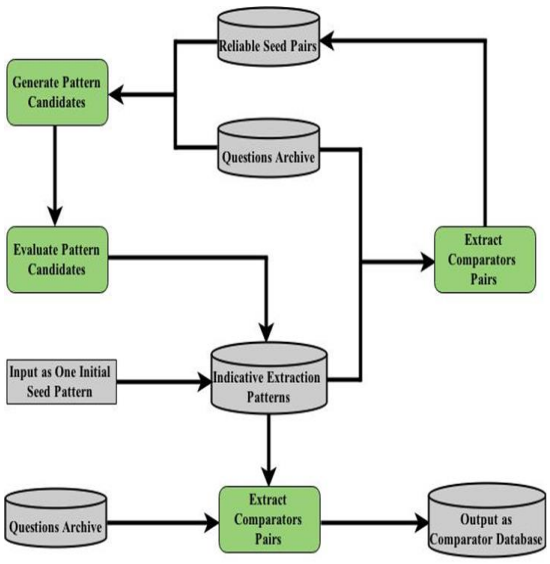

Fig.1. Flow chart of the bootstrapping algorithm

B. Proposed Work

The contribution of this paper is threefold, in first the extraction of comparative questions from questions achieved by users is investigated, in second, how to identify comparator aliases such as HCL and Hindustan Computer Limited and in third, how to separate ambiguous entities such "Brooklyn versus India" as location and "Brooklyn versus Nicole as celebrity is studied.

For extracting comparative questions based on user's questions archived online, we adopted a bootstrapping algorithm presented in [3].

The proposed indicative extraction pattern (IEP) which is weakly supervised is based on two key terms

- If given sequential pattern is able to extract numerous comparator pairs, it is highly probable to be an IEP.
- If a comparator pair is extracted by an IEP, the pair is considered as reliable one.

Algorithm1: Weakly Supervised Method pseudo code

Input: *CP, G*
Initialize solution: *Q ← {}, P ← {} Pnew ← {}*
CP + new ← {}
Repeat
P ← P + Pnew
Qnew ← compartiveQuestionidentify (CPnew)
Q ← Q + Qnew
For *qi ∈ G do*
If *ismatchexistingpatterns(p, qi)*
then
Q ← Q – qi
End if
End for
Pi ← mineGoodpatterns(Q)
cpnew ← { }
For *qi ∈ G*
do
cp ← extractcomparablerpatterns (p, qi)
If *cp ≠ NULL and cp ∉ CP*
then
CPnew ← CPnew + {CP}
End if
End for
Until *Pnew= {}*
Return P

This method inspires to work on sequential patterns which can recognize related comparative question and can mine comparators significantly. Bootstrapping works with single IEP at starting, from which comparator pair is extracted and considered as initial seed comparator pair. Afterword all questions containing these comparators are considered as comparative question. This method of extraction is iterative and self-learning hence called as Weakly Supervised Method i.e. needs less supervision.

C. Comparator Ranking

The comparability and graph based methods are examined rank possible comparators for user's input [1], [3], [18] which are described below,

1. Comparability-based ranking method :

Frequent comparison of entity with particular entity would make comparator more interesting. Based on this intuition, a simple ranking function R_{freq} *(c;e)* ranks comparators on the basis of number of times that a comparator c is get compared to the user's input e in online comparative question archive *Q*.

$$R_{freq}(c;e) = N(Q_{c,e}) \quad eq.\ (1)$$

Where (Q_c, e) is a set of questions from which comparator *c* and user input e can be extracted as a comparator pair. This method also known as frequency based Method. R_{rel} is another ranking function by combining reliability scores estimated in comparator mining phase

$$R_{rel}(c;e) = \sum_{q \in Q_{c,e}} R(p_{q,c,e}) \quad eq.\ (2)$$

Where *p, q, c, e* are pattern that is selected to mine comparator pair of c and e from question [3].

1. Graph -based ranking method :

Frequency is consider as efficient parameter for comparator ranking but the frequency-based ranking method [3] can suffer when an user input occurs rarely in collection of questions; for example, consider all possible comparators to the input are compared only once in questions. In such case, this method may get fail to results correct ranking result. Hence in addition to it representing ability should also be considered. We regard a comparator representative in comparison, if it is frequently used as a baseline or benchmark while making comparison of interested entity.

Graph based page rank technique is one of the solutions to get ability. A comparator can be considered as valuable comparator in ranking if it is compared to many other important comparators including the input entity. Based on this idea, Page Rank algorithm is examined to make ranking of comparators for a given input entity, which combine frequency and represent ability [3].

IV. Operating Envornment

This section enlists different hardware and software tools employed in developing decision making system, as follows

1. Hardware environment
   - Processor - Core i5 or higher version
   - RAM – 2 GB or more
   - Hard Disk – 100 GB or more
2. Software environment -
   - Operating System – Windows 8.1
   - Technology - Java and J2EE
   - Web Technologies - Html, JavaScript and CSS
   - IDE - Eclipse
   - Web Server - Tomcat
   - Database - My SQL

V. Results

In this section the partial results of decision making system developed in Java and MySQL for ambiguous entity extraction is presented. After sign in, the user will enter ambiguous entity such as "Brooklyn" which is city as well as the name of famous American actress. Fig. 2 shows

the screen shot of user's question where user can enter question and keywords.Fig.3 shows outcome for users question containing ambiguous entities description which standing for both city and celebrity

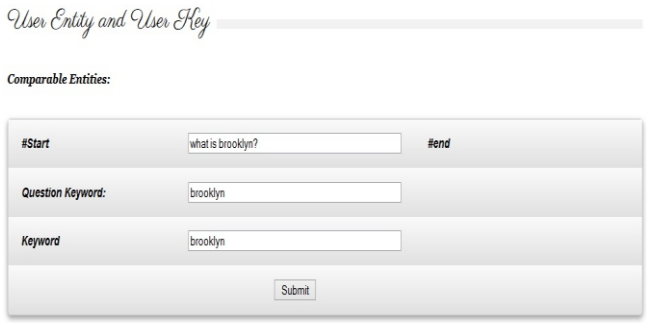

Fig. 2. User's input to the system.

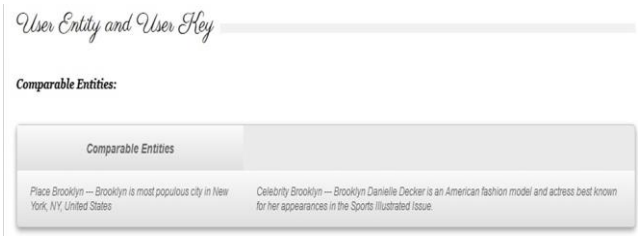

Fig. 3. Outcome for User's ambiguous entity input.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, we present a novel way to identify comparative questions and simultaneously extract their comparator pairs using weakly supervised technique. Bootstrapping extraction and identification process resides on the key sight that a good comparative question extracts good comparators likewise good comparators are present in good comparative questions. Also the method to identify ambiguous entities which is tricky in decision making process is presented. Experimental results for ambiguous and unambiguous entity are discussed. In future we would like to focus on rare extraction patterns more efficiently and precisely.

ACKNOWLEDGMENTS

The authors would like to thanks Dr. M. S. Nagmode, Principal MIT College of Engineering, Kothrud, Pune and Prof. A. S. Hiwale, HOD, IT department for their valuable suggestions and motivation for this research.

REFERENCES

[1] Nitin Jindal and Bing Liu, 'Identifying Comparative Sentences in Text Documents', Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 244-251, 2006.

[2] Li Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li, 'Comparable Entity Mining from Comparative Questions', Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10), 2010.

[3] Li Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li, 'Comparable Entity Mining from Comparative Questions', Knowledge and Data Engineering, IEEE Transactions on 25, no.7, pp. 1498-1509, 2013.

[4] LiShasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li, 'Comparative Entity Mining', U.S. Patent no. 8, 484, 201, July 2013.

[5] Narayane Shrutika and Sudipta Giri. 'A Review on Comparable Entity Mining', International Journal of Innovative Research in Computer and Communication Engineering , vol. 2, no. 12, pp. 7252-7257, 2014.

[6] Califf M. Elaine and Raymond J. Mooney, 'Relational Learning of Pattern Match Rules for Information Extraction', Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI '99/IAAI'99), pp. 328-334, 1999.

[7] Mooney J. Raymond and RazvanBunescu, 'Mining Knowledge from Text Using Information Extraction', ACM SIGKDD explorations newsletter 7.1, pp. 3-10, 2005.

[8] Cardie Claire, 'Empirical Methods in Information Extraction', Artificial Intelligence Magazine, vol. 18, pp. 65-79, 1997.

[9] Riloff Ellen and Rosie Jones, 'Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping', Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI '99/IAAI '99), pp. 474-479, 1999.

[10] Califf M. Elaine and Raymond J. Mooney, 'Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction', Journal of Machine Learning Research, vol 4, pp. 177–210, 2003.

[11] Mooney J. Raymond and Loriene Roy, 'Content-based Book Recommending using Learning for Text Categorization', Proceedings of the 5th ACM Conference on Digital Libraries, pp. 195–204, 2000.

[12] FreitagDayne, 'Toward General-purpose Learning for Information Extraction', Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and COLING-98 (ACL/COLING-98), pp.404–408, 1998.

[13] StephenSoderland, 'Learning Information Extraction Rules for Semi-Structured and Free Text', Machine Learning, vol. 34, nos. 1-3, pp. 233-272, 1999.

[14] Chang Chia-Hui and Shao-Chen Lui, 'IEPAD: Information Extraction Based on Pattern Discovery', Proceedings of the 10th international conference on World Wide Web (WWW' 01), 2001.

[15] Carlson Andrew, Justin Betteridge, Richard C. Wang, Estevam R. HruschkaJr, and Tom M. Mitchell, 'Coupled Semi-supervised Learning for Information Extraction', Proceedings of the 3rd ACM international conference on Web search and data mining, pp. 101-110, 2010.

[16] Nitin Jindal and Bing Liu, 'Mining Comparative Sentences and Relations', Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '06), vol. 22, pp. 1331-1336. 2006.