International Journal of Current Research in Life Sciences
Vol. 4, No. 4, pp. 184-186, April, 2015

www.ijcrls.com

Full Length Review Article

Mining Comparators from Comparative Questions

Akash Saindanvise, *Laxmi Venkatraman Tejas Shelke and Varun Varia

K.K.W.I.E.E.R, Nashik, India

Accepted 05th March, 2015; Published Online 30th April, 2015

ABSTRACT

One of the essential parts of human life is to compare one thing with another in order to take proper decisions. But it is difficult to find what parameters to compare and what could be the alternatives for it. To solve this difficulty we present a novel way to mine comparable entities from comparative questions. To ensure that accuracy is maintained we develop a weakly supervised bootstrap method. Experimentation has shown that this method has achieved accuracy of about 82.5% in comparative question identification and 83.3% in extraction of comparable entities.The results are far better than the state of art system that exists.

Key words: Bootstrapping, Comparative Questions, Comparators.

INTRODUCTION

An essential part of human life involves decision making. Take an example, if a person is interested in a certain product such as digital camera he would want to know the alternatives that are present before purchasing the product. This is a common task in our daily life but needs detailed knowledge and skill. PC Magazine and consumer report are amongst the few examples. A comparative activity typically involves: search for relevant web pages from the World Wide Web which have information about the target products, finding competing products, read reviews and to recognize its cons and pros. It is not easy to decide if two products can be compared or not as for various reasons two fruits like apple and orange cannot be compared. The situation can get more complicated if the entity has several functionalities. So to simplify things, we define comparative questions and comparators as:

Comparative question: The question which compares two or more things or entities is called comparative question.

Comparator: In a comparative question it is an entity which is target of question.

Related work

For finding related things if there should be an occurrence of an element, our work is like take a shot at recommender frameworks, which prescribes things to a specific client. A recommender framework is chiefly subject to the comparative gimmicks that are exhibit between the things.

Corresponding author: Laxmi Venkatraman,
K.K.W.I.E.E.R, Nashik, India.

We can take the case of Amazon which prescribes well known items, most offering items in light of the purchasing history of the client/client. Anyhow we can't say that prescribing an item is like contrasting a thing. If there should be an occurrence of Amazon, the motivation behind proposal is that the client ought to add more things to his truck and to enhance the offer of their products by proposing clients related or comparative things. While if there should arise an occurrence of correlation, we might want to help client to discover diverse options which would help them take better choices among the distinctive things which are to be looked at. For example, it is alright to prescribe "iPod speakers" or "iPod batteries" to a man/client who is keen on "iPod" yet we can't measure up these things with the iPod. However the thing which can be contrasted and "iPod" can be "iPhone" and the "PSP" which are in view of the likenesses of things. Despite the fact that they are all fit for playing music.

An "iPhone" is principally a cellular telephone, and "PSP" is mostly a versatile gaming gadget they are comparable in a few viewpoints yet they contrast those there is a need to analyse these items. Henceforth it is clear that comparator mining and thing proposal are connected however not same. The work which we have done on comparator mining is related to research on entity and relation extraction in information. In this field the most relevant work has been done by Jindal and Liu. Their method implied Class Sequential Rule (CSR) and Label Sequential Rule (LSR) which were used to identify comparative sentences and extract comparative relations respectively. But the problem with this method is that we can achieve high precision but have low recall. To solve this problem that has occurred we develop a weakly supervised bootstrapping pattern learning method.

Jindal and Liu 2006

Here we will provide a short summary of comparative mining method that was proposed by Jindal and Liu which acts as a base for comparison and is state-of-art work in this area.

CSR and LSR

CSR acts as a rule for classification. It is used for mapping the sequence pattern S(S1,S2.....Sn) to a class C. In this case the C could be either comparative or non-comparative. CSR is always associated with two parameters: support and confidence. The Support is the portion of the sequence in the collection containing S as a subsequence and Confidence is portion of sequence labelled as C in the sequence containing the S. These parameters are important to check if a CSR is reliable or not. LSR acts as a labelling rule. It maps the sequence of input pattern S (s1s2....si....sn) to a labelled sequence S'(s1s2...li ...sn) by replacing one token (si) in input with a particular label (li). This token is called as an anchor. The anchor in a corresponding sequence can be extracted if its corresponding label in a labelled sequence is what we need. LSR too is mined from an annotation corpus, hence each LSR has parameter: Support and Confidence, which are defined similarly as present in CSR.

Supervised Comparative Mining Method

Jindal and Liu treated relative sentence recognizable proof as an issue on grouping and similar connection extraction as an issue in view of extraction of data. Firstly they made physically a set of 83 pivotal words like beat, exceed and outflank which are likely markers of a comparison, these phrases were then utilized as turns to make grammatical form.
At the point when given a set of similar sentences, Jindal and Liu physically an noted two comparators one with mark $ES1 and other with name $ES2 and the gimmick are contrasted and mark $FT for each sentence. Jindal and Liu technique was connected just to thing and pronouns. To figure out the distinction between a thing and pronoun they included fourth name $NEF which is Non-Entity-Feature. These names were utilized with the marks #start (at the begin of sentence) and #end(at end of sentence) for era of sequenced information.
This method has following weaknesses:

- J&L's performance is dependent heavily on a set of comparative sentence indicative keywords.

- In many different ways the user can express a comparative sentence. To have a high recall, a large annotated training corpus is required.

Weakly Supervised Method for Mining Comparators

This technique is based on pattern, which has similar approach as that of Jindal and Liu's method, but it differs in a way where our approach uses sequential patterns to generate comparative questions and extracting comparators from them.
We begin our method by defining a sequential pattern as a sequence SP(sp1,sp2,…. sp n) where sp can be a word, a Part Of Speech (POS tag),start or end of a question etc. In this way, we define number of patterns. But, the pattern which allows us to identify the comparative question from a set of questions and ultimately generates comparators is said to be an Indicative Extraction Pattern(IEP). The IEP thus defined will be used as a reference and the questions will be compared with the pattern. If it matches with the pattern, then it can be classified as a comparative question and the corresponding tokens are listed out as comparators. Bootstrapping algorithm allows us to define a pattern instead of manually creating the set of keywords. This ensures high precision i.e. what percentages of retrieval documents are actually relevant to the query and recall i.e. what percentage of documents relevant to the query are retrieved. Let us consider an example to find out how the comparative question is defined and how the tokens are extracted. "<#start $C1/N or $C2/N? #end>" where, the question is defined between start and end tags. This question involves two comparators C1 and C2 which are nouns.

Comparable Entity Mining from Comparative Questions

Mining IEP

Now, we have to define the Indicative Extraction Pattern. For this, we will assume:

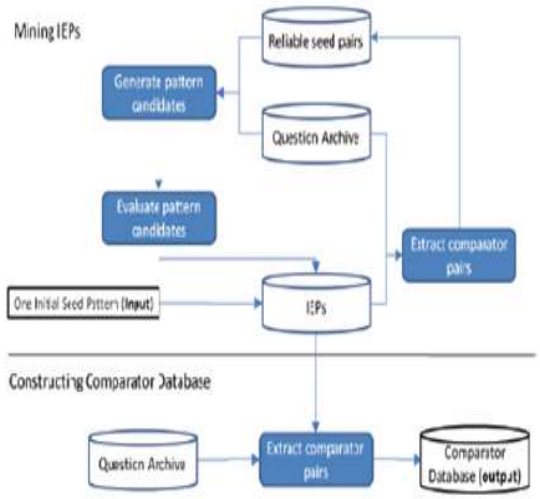
- If a pattern can be used to extract comparators, then it is an IEP.
- If the IEP can be used to extract the comparator pairs then it is reliable.

We design the bootstrapping algorithm by defining a single IEP. We will extract the comparators from this pattern. For every question in the question archive, we will check if that question has the pair, if yes then it is considered as a comparative question. Such questions will be then added to the IEP and the newly generated comparator pairs are extracted from them for further use. All the patterns are then evaluated for their reliability score. This process will continue until no new pattern is found. During this process, we have to carry out two important steps:

1. Pattern Generation
2. Pattern Evaluation

Pattern Generation

Assume that a given question is comparative. We have to generate pattern from this question, for this we will replace the comparators with symbols and add two keywords START and END at the beginning and end of the sentence respectively.



The pattern thus generated can be categorized as:

1. Lexical pattern
2. Generalized pattern
3. Specialized pattern

Lexical Pattern: This is a sequential pattern which consists of only words and comparators. It contains more than one comparator.

Generalized Pattern: This is a sequential pattern obtained by replacing the words with the parts of speech like noun, verb etc. Generalized patterns thus obtained are free of comparators.

Specialized Pattern: Sometimes a pattern can be too general. There can be many questions which are not comparative but still they match some pattern. To avoid this, we add POS tag to the comparator. We need to note that the final pattern generated is a mixture of all the above mentioned patterns namely lexical, generalized sand specialized patterns.

Pattern Evaluation

A reliability score $R\ (pi)$ for a candidate pattern piat iteration k can be defined as follows:

$$(pi)=\frac{\sum\forall cpj\in CP\wedge(k-1)\ NQ(\text{pi}\square\text{cpj})}{NQ\ (pi\rightarrow *)}$$

where,

pi can be used to extract comparator pairs.

$CP\wedge\ (k-1)$ indicates the comparator pairs stored until the $(k-1)t$ iteration
(x) means the number of questions that satisfy a condition x.
$pi\rightarrow\square$denotes a question that contains pattern.

But, the above formula lacks complete knowledge of the comparator pairs, since in the first iteration we take into account very few pairs. To overcome this drawback, we make use of look ahead procedure and the look ahead reliability is calculated as follows:

$$(pi)=\frac{\sum\forall cpj\in CP\text{rel}\wedge(k)NQ(\text{pi}\square\text{cpj})}{NQ\ (pi\rightarrow *)}$$

where (k) indicates a set of likely-reliable pairs.
We now combine the above two formulas to calculate the final reliability score (pi)

$$k=\lambda\cdot Rkpi+(1-\lambda)\cdot(pi).$$

Conclusion

In this paper, we have worked on to identify comparative questions and to mine comparator pairs from them. For this, we have used Bootstrapping algorithm which differs from J & L method in a manner where we define sequential patterns in contrast with the conventional keyword method. For this, we have prepared a set of questionnaires that we have obtained from the websites which are commonly visited by the users.
After comparing the conventional and modern methods, we have come to a conclusion that our method is more effective in improving recall and precision. Besides this, we have further improved our method in which we have designed a mechanism to identify aliases such as distinguishing between the acronyms and their corresponding long forms. We have also developed a method to identify the ambiguous entities and put them into their proper categories based on what is being compared.

REFERENCES

Comparable Entity Mining from Comparative Questions, IEEE Transactions on Knowledge and Data Engineering, 2013.

Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In *Proceedings of AAAI '06*.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level Bootstrapping. In *Proceedings of AAAI '99 /IAAI '99*, pages 474–479.

****