

AUTOMATIC MINING OF COMPARABLE ENTITIES

Sharda Dabhekar¹, Prof. Shubhangi Vairagar²

¹ M.E. Student ,Computer Department, Siddhant College of Engineering, Maharashtra, India

² HOD, Computer Department, Siddhant College of Engineering , Maharashtra, India

ABSTRACT

Making Comparisons between things is a typical part of human decision making process. But however, it is difficult to know what are to be compared and what can be the alternatives. For eg. if someone is interested in certain products such as digital cameras, then he /she would want to know what the alternatives are and compare different cameras before making any purchase. This type of comparison activity is very common in our daily life but requires high knowledge skill in order to make much better choice. Therefore, to address this difficulty, we are presenting a novel way to automatically mine comparable entities from comparative questions that users posted online. The literature review revealed that this method gets an F1-measure of 82.5 percent in identification of comparative question and 83.3 percent in comparable entity extraction.

Keyword : - Entity comparing, Attribute extraction, decision making, bootstrapping, Question identification.

1. INTRODUCTION

For making decisions we normally Compare alternative options. If someone is interested in certain products such as digital cameras, he or she would want to know what the alternatives are and compare different cameras before making a purchase. This type of comparison activity is very common in our daily life but requires high knowledge skill. In the World Wide Web era, a comparison activity typically involves: search for relevant web pages containing information about the targeted products, find competing products, read reviews, and identify pros and cons. In this paper, we focus on finding a set of comparable entities given user's input entity to mine comparators from comparative questions, we first have to detect whether a question is comparative or not. According to our definition, a comparative question has to be a question with intent to compare at least two entities. Please note that a question containing at least two entities is not a comparative question if it does not have comparison intent. However, we observe that a question is very likely to be a comparative question if it contains at least two entities. We leverage this insight and develop a weakly supervised bootstrapping method to identify comparative questions and extract comparators simultaneously. The comparative questions and comparators can be thus defined as:

e.g. " Which to purchase, iPod or iPhone?" We call "iPod" and "iPhone" in this illustration as comparators. In this paper, we characterize near inquiries and comparators as:

- Comparative inquiry: An inquiry that plans to look at two or more elements and it needs to say these elements unequivocally in the inquiry.
- Comparator: An element which is an objective of correlation in a near inquiry. By definitions, Q1 and Q2 underneath are not similar inquiries while Q3 is. "iPod Touch" and "Zune HD" are comparators.

Q1: "Which one is better?"

Q2: "Is Lumix GH-1 the best camera?"

Q3: "What's the contrast between "iPod Touch" and "Zune HD?"

comparator mining is related to the research on entity and relation extraction in information extraction. Specifically, the most relevant work is mining comparative sentences and relations. Their methods applied class sequential rules (CSR) and label sequential rules (LSR) learned from annotated corpora to identify comparative sentences and extract comparative relations respectively in the news and review domains. The same techniques can be applied to comparative question identification and comparator mining from questions. This method typically can achieve high precision but suffer from low recall. Cannot find easily to mining the data. We present a novel weakly supervised method to identify comparative questions and extract comparator pairs simultaneously. We rely on the key insight that a good comparative question identification pattern should extract good comparators, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. By leveraging large amount of unlabeled data and the bootstrapping process with slight supervision to determine four parameters. To ensure high precision and high recall, we develop a weakly-supervised bootstrapping method for comparative question identification and comparable entity extraction by leveraging a large amount of data. Pattern Generation (comparable Entity) has three different forms:

- **Lexical patterns:**

A lexical pattern can be too specific. Thus, we generalize lexical patterns by replacing one or more words with their POS tags. $2n-1$ generalized patterns can be produced from a lexical pattern containing N words excluding $\$C$ s.

- **Specialized Patterns:**

In some cases, a pattern can be too general. For example, although a question "ipod or zune?" is comparative, the pattern "<\$C or \$C>" is too general, and there can be many non-comparative questions matching the pattern, for instance, "true or false?". For this reason, we perform pattern specialization by adding POS tags to all comparator slots. For example, from the lexical pattern "<\$C or \$C>" and the question "ipod or zune?", "<\$C/NN or \$C/NN?>" will be produced as a specialized pattern.

- **Pattern Evaluation (Comparable Questions):**

In complete knowledge about reliable comparator pairs. For example, very few reliable pairs are generally discovered in early stage of bootstrapping. In this case, the value of might be underestimated which could affect the effectiveness of on distinguishing IEPs from non-reliable patterns. We mitigate this problem by a look ahead procedure. Let us denote the set of candidate patterns at the iteration k by C_k . We define the support S for comparator pair c which can be extracted by C_k and does not exist in the current reliable set.

1.1 Information extraction

Information extraction is the task of automatically extracting structured information from unstructured readable documents. For purchasing a product a wealth of information that can be very helpful in accessing the comparable entities and opinions toward products. Almost every day people are faced with a situation that must decide upon one thing or the other. To make better decisions probably attempt to compare entities that the customer are interesting in. These days many web search engines are helping people look for their interesting entities. Therefore a comparison mining system, which can automatically provide a summary of comparisons between two entities from a large quantity of web documents, would be very useful in many areas such as marketing. The work is divided into two tasks to effectively build a comparison mining system. First classify the sentences into comparatives and non-comparatives and the second is related to comparative mining.

1.2 Related work

If consideration of entity determination is done, the proposed systems are similar to recommender systems that recommend various items to the users. Item similarity or correlation between the various user log are the basis of most of the recommendation systems (Linden et al., 2003). Like, Amazon E commerce system recommends only those products to its customers which are relevant to the previous purchase logs of the user. But unfortunately recommending some product to the user is not equivalent to comparator or entity identification. Amazon's major aim in recommending the similar products is because it wants users to add more products to their shopping cart and increase their business. On the other hand, comparison between entities make users come to a proper decision regarding which comparator or entity to be purchased. Despite the fact that they are all music players, "iPhone" is basically a cellular telephone, and "PSP" is for the most part a compact amusement gadget. They are comparative

additionally distinctive in this manner ask examination with one another. It is clear that comparator mining and thing suggestion are connected yet not the same. In this manner the proposed framework concentrates on element extraction from data extraction process.

2. PROPOSED SYSTEM

The proposed weakly supervised technique is a pattern-based approach same as like of J&L's method, but both techniques differ in various aspects: The proposed system makes use of sequential pattern that identifies comparative question and extract comparators simultaneously, instead of the existing techniques which make use of separate CSR's and LSRs:

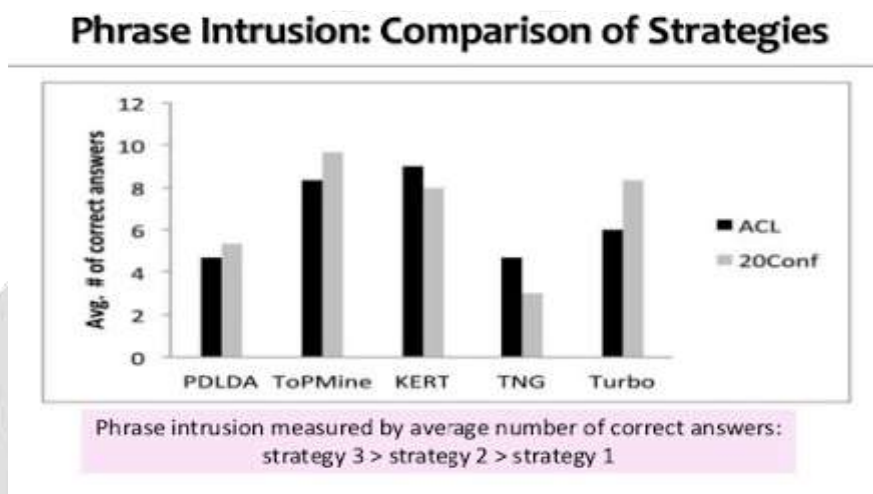


Chart -1: Comparison of Strategies

In our approach, a sequential pattern is defined as a sequence $S(s_1s_2...s_i...s_n)$ where s_i can be a word, a POS tag, or a symbol denoting either a comparator ($\$C$), or the beginning ($\#start$) or the end of a question ($\#end$). A sequential pattern is called an indicative extraction pattern (IEP) if it can be used to identify comparative questions and extract comparators in them with high reliability. If a match of the question is found with an indicative extraction pattern, it is considered as a comparative question and then the token sequences related to the comparator slots in the indicative extraction pattern are further extracted as comparator entities. If a question matches the large number of indicative extraction pattern, the longest indicative extraction pattern is used as final IEP. Thus we avoid creating a manual set of keywords and create an automatic set of IEPs. The proposed system demonstrates the automatic IEP generation using the bootstrapping technique with minimal supervision and thereby taking advantage of a huge unlabeled question set. Table 1 shows an examples of one such sequential pattern. The proposed system also allows POS constraint on comparator entities as shown in the pattern " $< \$C/NN$ or $\$C/NN? \#end$ ". It indicates that a valid comparator entity must have a NN POS tag.

A. Working of Proposed System

The proposed system flow is as mentioned below:

- The Raw data is used as input as a text from which the segmentation of sentences and IEPs are carried out."
- Sentence segmentation is followed by data loading where in *list of strings* are fetched out from the loaded data.
- Post Segmentation, the tokenization process of the *list of strings* is carried out for creating the various tokens to generate and *list of list of strings*.

- Once the tokenization process is completed, the parts of speech is tagged to the tokens generated in previous step i.e. *list of list of strings* are used from tokenized sentences to form the post tagged sentences i.e. *list of tuples*.
- On post tagging, the list of list of tuples are used as post tagged sentences for entity determination process. Entity detection comprised of chunk creation for the tagged words from the *list of list of tuples*.
- Post chunking, once the entities are identified, the relation between the entities are identified for determining actual relation between the entities and its existence in the sentence.
- The identified relations are nothing but the actual output of the proposed system which is determined from the raw data taken as input.

B .Mining Indicative Extraction Patterns

The proposed weakly supervised mining technique using IEP is based on two vital assumptions:

- If any sequential pattern extracts multiple comparator pairs, it is more important to be an IEP.
- If an IEP helps in extracting a comparator pair, the pair is considered to be reliable

Calculation Based on aforementioned two presumptions, the bootstrapping calculation is planned as said. as appeared in Figure. underneath. Introduction of bootstrapping calculation is with a solitary IEP and seed comparator sets are separated from it. At that point all the inquiry from the accumulation set are recovered and promote viewed as near inquiries if the inquiry coordinates the seed pair. Every single consecutive example which are conceivable are produced and further assessed by mining their unwavering quality score, from the near inquiries and comparator pair. Designs which are included into IEP storehouse are really assessed as the genuine solid sets.

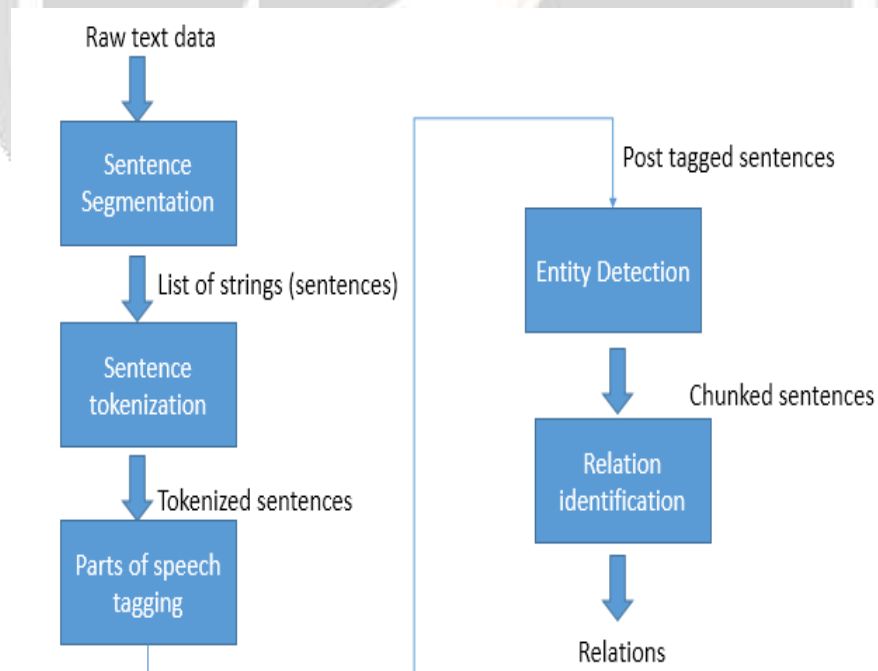


Fig -1: Flow of the Proposed System

2.1 Comparative Study

Most of the earlier designed systems did not extract the comparable entity from a comparable question. All the questions were considered as the question to be considered. The proposed system distinguishes between the Comparable question and the other non-comparable questions. The proposed and existing systems identified between questions as mentioned in the table below:

Table -1: Comparative analysis of R question identification

Sr.No	Input Questions		
	Questions	Existing system	Proposed System
1	Is mobile Nokia lumia 725 good ? ¹	✓	✗
2	Nokia is better or Samsung ?	✓	✓
3	Lithium battery is good or not?	✓	✗
4	Apple smapyphone is better or Samsung?	✓	✓

2.2 Examples of Comparator Extraction

By applying our bootstrapping method to the entire source data (60M questions), 328,364 unique comparator pairs were extracted from 679,909 automatically identified comparative questions. Lists frequently compared entities for a target item, such as Chanel, Gap, in our question archive. As shown in the table list, our comparator mining method successfully discovers realistic comparators. Our comparator mining results can be used for a commerce search or product recommendation system. For example, automatic suggestion of comparable entities can assist users in their comparison activities before making their purchase decisions. Also, our results can provide useful information to companies which want to identify their competitors.

.	Chane	Gap	iPod	Kobe	Canon
1	Dior	Old Navy	Zune	Lebron	Nikon
2	Louis	Amerian	Mp3	Jordan	Sony
3	Coach	Banana	PSP	MJ	Kodak
4	Gucci	Guess	Cell	Shaq	Panasonic
5	Prada	ACP	iPhone	Wade	Casio
6	Lancom	Old	Creative	T-mac	Olympus
7	Versace	Hollister	Zen	Lebron	Hp

List of Examples of comparators for different entities

For example, for Chanel, most results are high end fashion brands such as Dior or Louis Vuitton, while the ranking results for Gap usually contains similar apparel brands for young people, such as Old Navy or Banana Republic. For the basketball player Kobe_, most of the top ranked comparators are also famous basketball players. Some interesting comparators are shown for Canon (the company name). It is famous for different kinds of its products, for example, digital cameras and printers, so it can be compared to different kinds of companies. For example, it is compared to HP Lexmark, or Xerox, the printer manufacturers, and also compared to Nikon, Sony, or Kodak, the

digital camera manufactures. Besides general entities such as a brand or company name, our method also found an interesting comparable entity for a specific item in the experiments. For example, our method recommends „Nikon d40i_„Canon rebel xti_„Canon rebel xt_„Nikon d3000_„Pentax k100d_ Canon eos 1000d_ as kon 40d

3. PATTERN GENERATION

We now give the precision, recall and F-score results. All the results were obtained through 5-fold cross validations.

Identifying gradable comparatives: NB using CSRs and manual rules as the attribute set gave a precision of 82% and a recall of 81% (F-score = 81%) for identification of gradable comparative sentences. We also tried various other techniques, e.g., SVM (Joachims 1999), CSR rules only, etc., but the results were all poorer. Due to space limitations, we are unable to give all the details. (Jindal & Liu 2006) has all the results using a larger dataset.

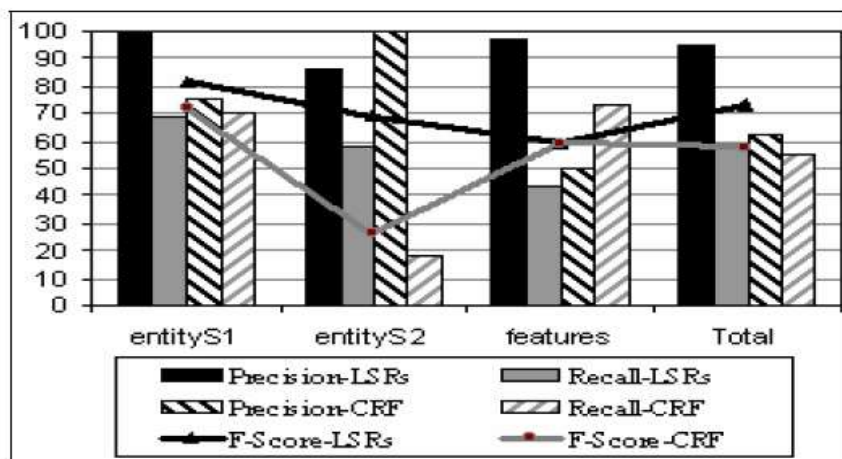


Chart -2: Precision-Recall and F-Score results of LSRs and CRF for extracting relation entries

3.1 Indicative Extraction Pattern (IEP)

In the proposed situation, S(s1s2 ... si ... sn) goes about as a consecutive example, where si can be a POS tag, a word, or an image speaking to either a starting (#start) or a comparator (\$C), or the end of an inquiry (#end). A consecutive additionally meant as (IEP) demonstrative extraction design, which can be utilized to decide relative inquiries and concentrate the sentence tokens with greatest unwavering quality.

Sequential Patterns
<#start which city is better, \$C or \$C? #end>
<, \$C or \$C? #end>
<#start \$C/NN or \$C/NN? #end>
<which NN is better, \$C or \$C?>
<which city is JJR, \$C or \$C?>
<which NN is JJR, \$C or \$C?>

Fig -2: Candidate Indicative Extraction Pattern (IEP) Example

If a match of the question is found with an indicative extraction pattern, it is considered as a comparative question and then the token sequences related to the comparator slots in the indicative extraction pattern are further extracted

as comparator entities. If a question matches the large number of indicative extraction pattern, the longest indicative extraction pattern is used as final IEP. Thus we avoid creating a manual set of keywords and create an automatic set of IEPs. The proposed system demonstrates the automatic IEP generation using the bootstrapping technique with minimal supervision and thereby taking advantage of a huge unlabeled question set. Table 1 shows an examples of one such sequential pattern. The proposed system also allows POS constraint on comparator entities as shown in the pattern “<, \$C/NN or \$C/NN? #end>”. It indicates that a valid comparator entity must have a NN POS tag

4. CONCLUSION

In this paper, proposed framework executed a novel feebly managed procedure to decide near inquiries and concentrate the solid comparator matches all the while. The proposed framework depends on the key presumption that a decent near inquiry determination example ought to have the capacity to concentrate great comparators, and a decent comparator pair must happen in a decent similar inquiry to bootstrap the extraction and determination process. By utilizing immense measure of unlabeled information and the proposed bootstrapping process with slight measure of supervision to decide parameters, the proposed framework can separate the solid comparators from the data crude information.

5. ACKNOWLEDGEMENT


I would like to express my sincere thanks to my Project guide Prof. Shubhangi Vairagar mam, Head of Department, Siddhant college of Engineering for his motivation and valuable suggestions which truly helped me in improving the quality and effectiveness of this paper. I take this opportunity to express my thanks to my teacher, family and friends for their encouragement and support and valuable suggestions.

6. REFERENCES

- [1] G. Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In Proceedings of AAAI'99/IAAI'99.
- [2] Claire Cardie. 1997. Empirical methods in information extraction. AI magazine, 18:65–79.
- [3] Dan Gusfield. 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, New York, NY, USA.
- [4] Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In Proceedings of WWW '02, pages 517–526.
- [5] Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In Proceedings of WWW '03, pages 271–279.
- [6] Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In Proceedings of SIGIR '06, pages 244–251.
- [7] Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In Proceedings of AAAI '06.
- [8] Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT, pages 1048–1056.
- [9] Greg Linden, Brent Smith and Jeremy York. 2003. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, pages 76-80.
- [10] Raymond J. Mooney and Razvan Bunescu. 2005. Mining knowledge from text using information extraction. ACM SIGKDD Exploration Newsletter, 7(1):3–10.
- [11] Dragomir Radev, Weiguo Fan, Hong Qi, and Harris Wu and Amardeep Grewal. 2002. Probabilistic question answering on the web. Journal of the American Society for Information Science and Technology, pages 408–419.
- [12] Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In Proceedings of ACL '02, pages 41–47.
- [13] Stephen Soderland. 1999. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233–272.
- [14] Veena G R Kumar received the B.E degree from Ponjesly college of Engineering, Nagercoil in 2008 and currently doing M.E in VINS christian college of Engineering, Nagercoil. Her area of interest is in data mining and networking.

- [15] K. Kundgir, P. Dere, S. Mohite, P. Mithari, "A Study on Comparable Entity Mining from Comparative Questions", *International Journal of Advancement in Engineering Technology, Management and Applied Science*, vol. 1, no. 2349-3224, 2014, 62-66.
- [16] S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li, "Comparable Entity Mining from Comparative Questions," *IEEE Transactions On Knowledge And Data Engineering*, vol. 25, no. 7, 2013, 1498-1509.
- [17] S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li, "Comparable Entity Mining from Comparative Questions," *Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '10)*, 2010.
- [18] 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

BIOGRAPHIES

	<p>Sharda Dabhekar obtained her bachelor Degree in Computer Engineering from university of RTMNU in 2012. At present she is pursuing M.E. in Computer Engineering from Department of Computer Engineering Siddhant College of Engineering, Sudumbare, Pune. Maharashtra, India.</p>
	<p>Shubhangi Vairagar Siddhant College of Engineering, Pune, Maharashtra, India. She received the B.E.,M.E. & PHD degree in Computer Science And Engineering, Pune, India.</p>