**Comparable Entity Mining from Comparative Questions**

**LITERATURE SURVEY**

Identifying Comparative Sentences in Text Documents.

N. Jindal and B. Liu, “Identifying Comparative Sentences in Text Documents,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), 2006, 244-251.

Comparisons can be subjective or objective. It’s more like comparing one object with another, for example, this city A is much better than city B. This paper studies how to identify comparative sentences. The paper clearly addresses the problem of identifying the comparative sentences. The approach first categorizes comparative sentences into different types and then presents integrated pattern discovery in identifying comparative sentences from text documents. The classification of sentences into comparative and non-comparative is the basic idea of the paper.

The approach takes into account of the Class sequential rules in building the learning model. The basic approach followed is the combination of the class sequential rules and machine learning. Use of machine learning methods to overcome the previous drawbacks is used in distinguishing comparative and non – comparatives. The paper has taken into account of the user preferences and implicit comparators also. The way of extracting comparative sentences from the text based is used in many applications. The main disadvantage is that it suffered from low recall though high precision was attained.

Mining Comparative Sentences and Relations

N. Jindal and B. Liu, “Mining Comparative Sentences and Relations,” Proc. 21st Nat’l Conf. Artificial Intelligence (AAAI ’06), 2006.

The basic idea of the paper is to have comparative relations from the already identified comparative questions. One way to evaluate an event is to compare with similar event. Class sequential rules and label sequential rules are used to accomplish the above task. From the input, they identify the comparative sentences and classify the identified comparative sentences into different classes. After that extract comparative relations which involves extraction of entities and their features that are being compared and comparative keyword. The main idea of the paper hovers around the basic agenda, the way of evaluation is to compare with similar entity. It suffered from low recall, though the precision is reasonably high.

Mining Knowledge from Text using Information Extraction.

R.J. Mooney and R. Bunescu, “Mining Knowledge from Text Using Information Extraction,” ACM SIGKDD Exploration Newsletter, vol. 7, no. 1, 2005, 3-10

[13] M. E. Califf and R. J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. Journal of Machine Learning Research, 4:177– 210, 2003.

[28] C. D. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.

[59] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In Proceedings of the Third Workshop on Very Large Corpora, 1995.

Paper discusses about use of two approaches in natural language information extraction for text mining. The first is extraction of general knowledge directly form the text and the second one is to first extract the structured data from the unstructured data or the semi structured data by applying traditional knowledge discovery in database methods. Supervised machine learning methods were preferred than the manually developing patterns as the machine learning method is successful in developing robust information extraction systems. Many IE systems simply treat text as sequence of tokens however many others treat them other tools like POS, learned extraction rules [13], phrase chunkers to identify phrases to extract [59], others use lexical semantic databases, as WordNet [28].

The whole paper revolves around extraction of information which is structured in nature from unstructured data or semi structured data. The main disadvantage is that it cannot meet the demand of the ever increasing corpus.

Learning Surface Text Patterns for a Question Answering System

D. Ravichandran and E. Hovy, “Learning Surface Text Patterns for a Question Answering System,” Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL ’02), pp. 41-47, 2002.

Open domain question and answering systems use external knowledge and tools for answering. A fairly list of surface patterns was a way to evaluate them. Certain type of answers are expressed using characteristic phrases, authors try to automatically acquire patterns and try to measure the accuracy of the patterns. Pattern learning algorithm is described by constructing a pattern for each individual question type.

The algorithm is to select an initial example and search for it in a search engine and select the top thousand results and apply a sentence breaker to the selected documents. In the selected documents retain only those sentences that contain both the question and answer term, then remove the tags, tokenize the input and pass the retained sentence through suffix tree constructor. Lastly pass each phrase through filter and retain only those phrases / sentences which contain both question and answer. The proposed system assumes that each sentence is a simple sequence of words. The disadvantage of the system is it doesn’t make a clear distinction between lower and upper case.

**Relational Learning of Pattern-Match Rules for Information Extraction**

M.E. Califf and R.J. Mooney, “Relational Learning of Pattern- Match Rules for Information Extraction,” Proc. 16th Nat’l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence (AAAI ’99/IAAI ’99), 1999

Information extraction systems are difficult and time consuming to build and they are domain specific and its complex to port to new domains. Robust Automate Production of Information Extraction system proposed, learns rules for the complete information extraction. The resulting rules extract the desired items directly form the text without parsing or processing. The learning algorithm primarily consists of specific to general search of patterns. First most specific patterns are created for each example. Pre filler contains item for each word from the beginning of the document to the word immediately preceding the filler. Now RAPIER attempts to compress the specific rules and try to generalize them, initially it take two rules and tries to create a general one. RAPIER system uses relational learning to construct unbounded pattern match rules for information extraction.

**Comparable Entity Mining from Comparative Questions**

[1] Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In Proceedings of SIGIR ’06, pages 244–251.

In the World Wide Web era comparison involves search of the term in the web, scan through the pages, find reviews and so on. The paper focuses on finding the alternatives to a word. The comparators are mined from the comparative questions. If some user is posting a question which has tries to find the relation between two words, then that means that the user is trying to find the relation between those two entities. We simply rely on this search from the user in finding comparative entities.

What can be compared and what cannot be compared, how to identify whether the question supplies is comparative question and how to retrieve the comparators without comprising on precision and recall is the main idea of the paper.

The authors use the boot strapping algorithm to overcome the drawback of the previous approach proposed by Jindal and Liu [1] which lags in low precision and recall in entity mining. Their approach is using Class sequential rules and Label sequential rules. The proposed system makes three kinds of patterns in evaluating whether a given user question are comparable or not. It uses Lexical patterns, Generalized patterns and Specialized patterns. The boot strapping algorithm process starts with a single IEP (indicative extraction pattern), from it set of initial seeds of comparator pairs are extracted. All the questions which contain these comparator pairs are considered as comparator questions. All the reliable ones are added to IEP repository. In the paper they have presented a weakly supervised method to identify comparative questions and extract comparators.