Mining Event-based Commonsense Knowledge from Web using NLP Techniques

Priya K V
Department of Computer Science
Karunya University, Coimbatore

Mathew Kurian
Department of Computer Science
Karunya University, Coimbatore

ABSTRACT
The real life intelligent applications such as agents, expert systems, dialog understanding systems, weather forecasting systems, robotics etc. mainly focus on commonsense knowledge And basically these works on the knowledgebase which contains large amount of commonsense knowledge. The main intention of this work is to create a commonsense knowledgebase by using an effective methodology to retrieve commonsense knowledge from large amount of web data. In order to achieve the best results, it makes use of different natural language processing techniques such as semantic role labeling, lexical and syntactic analysis.

Key Words: Automatic statistical semantic role tagger (ASSERT), lexico-syntactic pattern matching, semantic role labeling (SRL)

1. INTRODUCTION
Everyday peoples are coming to face a lot of situations or circumstances. In those circumstances, some of them are familiar and some of them are unfamiliar. In order to respond to those situations, a tremendous amount of knowledge is necessary. There are mainly two kinds of knowledge such as specialist's knowledge and commonsense knowledge. Specialist knowledge includes the knowledge possessed by mathematicians, engineers or scientists. But the commonsense knowledge is the knowledge possessed by every people, even small two year children possess.

Commonsense knowledge can be defined as the ability to analyze a situation based on its context using millions of common knowledge. Commonsense knowledge include basic facts about events and their responses, facts about beliefs and their desires or facts about data and how they obtained. This type of knowledge is attained by the process of living and growing in this world. Even a two year old child knows that if he drop a glass of water, the glass will break and water will spill on the podium. Or he knows that if he holds a knife by its blade, then the blade will cut his hand. So the human people can respond to different situations in various ways. Computers can be used in various applications in order to minimize the human effort. Then the computers are programmed with vast amount of details for this purpose. But capabilities of computers do not match the capabilities of human beings. Normally computers lack commonsense knowledge .If it is possible to give this commonsense to machines, then these machines can behave as a human.

The remaining sections are organized as follows: In Section 2 a detailed literature survey of related works is discussed. In Section3 the proposed methodology to extract the commonsense knowledge from the web searches results. Then the work is concluded and the future work to attain the entire objectives is briefed in Section 4.

2. RELATED WORK
In order to develop an efficient technique for the automatic retrieval of event-based commonsense knowledge from web, it is inevitable to study and analyze the related techniques and methods. There have been a significant number of studies attempting to automatically retrieve knowledge using text mining approaches. The purpose is to automatically find the relationship between concepts so that the process of building semantic resources can be fully or partially automated. Many of the studies retrieve knowledge from certain machine readable dictionaries. In order to increase the scope of coverage of the commonsense knowledge, many studies turned to use of more large scale free – text resources, especially the web.

A lexical knowledge base constructed automatically from the definitions and example sentences in two machine-readable dictionaries (MRDs), MindNet embodies several features that make a difference with MRDs. It is, however, more than this static resource alone. MindNet represents a general methodology for acquiring, structuring, accessing, and exploiting semantic information from natural language text [9]. MindNet is produced by a fully automatic process, based on the use of a broad-coverage NL parser and it is built regularly as part of a normal regression process. The main benefit of this system is that the problems introduced by daily changes to the underlying system or parsing grammar are quickly identified and fixed. Rather than using NLP, the automatic procedures such as MindNet's provide the only credible prospect for acquiring world knowledge on the scale needed to support common-sense reasoning. The broad coverage parser used in the Microsoft Word 97 grammar checker is similar to which is used in the extraction process in MindNet. This parser produces syntactic parse trees and deeper logical forms, to which rules are applied that generate corresponding structures of semantic relations. The parser has not been specially tuned to process dictionary definitions and all enhancements to the parser are moulded to handle the immense variety of general text, of which dictionary definitions are simply a modest subset.

The large network of inverted semrel structures are contained in MindNet. These inverted semrel structures facilitate the access to direct and indirect relationships between the root word of each structure, which is the headword for the MindNet entry containing it, and every other word contained in the structures. These relationships, consisting of one or more semantic relations connected together, constitute semrel paths between two words. Similarity and inference are the different methods used in

MindNet to identify the similarity between words. But some researchers have failed to distinguish between substitution similarity and general relatedness. This similarity function mainly focuses on measuring substitution similarity and a function is also used for producing clusters of generally related words. This similarity procedure is based on the top-ranked semrel paths between words. The main drawback of MindNet is that the detailed information from the parse, both morphological and syntactic, sharply reduces the range of senses that can be plausibly assigned to each word.

REES is a large-scale relation and event extraction system which extracts many types of relations and events with a minimum amount of effort, but high accuracy [10]. This can handle 100 types of relations and events and it does in a modular and scalable manner. A declarative lexico driven approach is used in this system and this approach requires a lexicon entry for each event-denoting word, which is generally a verb. The lexicon entry specifies the syntactic and semantic restrictions on the verb's arguments. Another application of commonsense knowledge is the MAKEBELIEVE system-it is a story generating agent which make use of commonsense knowledge for generating stories. The initial story seed are produced by the user and based on these inputs; it will create the fantastic stories [11]. For this it is needed to collect the ontology from the Open Mind Commonsense Knowledgebase. Binary causal relations are extracted from these input sentences and stored as crude trans-frames. By performing fuzzy, creativity-driven inference over these frames, creative "causal chains" are produced for use in story generation. This system has mostly local pair-wise constraints between steps in the story, though global constraints such as narrative structure are being added. And this system also makes use of structuralist and transoformalist approaches. But the ambiguity inherent in any natural language representation makes it difficult to resolve the bindings of agents to actions when more than one agent is involved. And this ambiguity precludes MAKEBELIEVE from being able to tell multiple character stories.

There are basically two large -scale commonsense knowledge base such as Lenat's CYC and Open Mind Commonsense (OMCS).CYC contains s over a million hand-crafted assertions, expressed in formal logic while OMCS has over 400,000 semi-structured English sentences, gathered through a web community of collaborators. Sentences in OMCS are semi-structured, due to the use of sentence templates in the acquisition of knowledge, so it is relatively easy to extract relations and arguments [14].

ConceptNet is a freely available commonsense knowledgebase and natural language processing toolkit which supports many practical textual-reasoning tasks over real-world documents including affect-sensing, analogy-making, and other context oriented inferences. This knowledgebase is a semantic network presently consisting of over 1.6 million assertions of commonsense knowledge encompassing the spatial, physical, social, temporal, and psychological aspects of everyday life [13]. Whereas similar large-scale semantic knowledge bases like Cyc [4] and WordNet [7] are carefully handcrafted, ConceptNet is generated automatically from the 700,000 sentences of the Open Mind Common Sense Project – a World Wide Web based collaboration with over 14,000 authors. ConceptNet is a unique resource which contains a wide range of commonsense concepts and relations, such as those found in the Cyc knowledgebase [6]. But it is structured not as a complex and intricate logical framework, but rather as a simple, easy-to-use semantic network, like WordNet. While ConceptNet still supports many of the same applications as WordNet, such as query expansion and determining semantic similarity, its focus on concepts-rather-than-words, it's more diverse relational ontology, and its emphasis on informal conceptual-connectedness over formal linguistic-rigor allow it to go beyond WordNet to make practical, context-oriented, commonsense inferences over real-world texts. The main drawback of this is to continue to make progress in textual-information management; vast amounts of semantic knowledge are needed to give this software the capacity for deeper and more meaningful understanding of text. And without additional insight into how a concept is generally interpreted by default (which would require a difficult, deep parse), it can only make heuristic approximations as to the relative contributions of the verb, noun phrase, attribute, and prepositional phrase to the meaning of a concept. It is quite difficult to produce useful stand-alone objective evaluations of knowledgebase quality. Computing conceptual similarity using lexical inferential distance is very difficult i.e. similarity scoring is not accurate.

3. EXPERIMENTAL METHODOLOGY

The main focus of this work is to develop a commonsense knowledge base effectively and efficiently. In order to create such a knowledge base from the mass amount of web data, enormous effort is required. This system mainly focuses to develop a methodology for retrieving the event-based commonsense knowledge from the web. For retrieving the event-based commonsense knowledge, the integration of different techniques such as lexico syntactic pattern matching and semantic role labeling is required [1] .After retrieving the knowledge items from the web, evaluate those results and create a commonsense knowledge base by adding those components. Then the users can easily retrieve the commonsense knowledge from this knowledge base. So this system mainly consists of four different modules. Each of them is briefly described in the following sections. The main modules include content extraction, semantic role identification, semantic role verification and knowledge distillation.

3.1 Content Extraction

The first step of this framework is to extract the raw sentences corresponding to the target knowledge item. For that an event is given to the web search engine like Google. Then the query will be formulated using lexico-syntactic pattern matching through web search engine. In order to find out the semantic relations, it will automatically do the lexical analysis and syntactical analysis. After that web search engine gives the response as a list of web pages or snippets. From each snippet or webpage, all the contents or sentences should be extracted. In the web search results most of the sentences belong to the dynamic modality. Dynamic modality means it describes a factual situation about the subject of the sentence [1].In order to extract the content of a web page which contains the required knowledge item; the first step is to create the web browser. And after entering the required URL in this web browser, the content of that particular web page

which pointed by the given URL will be extracted and it will be stored in a text file.

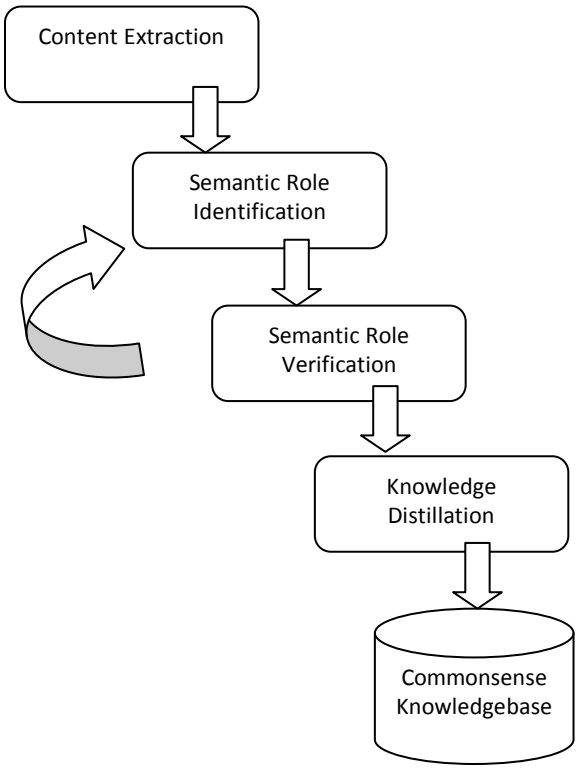

Fig.1 Framework for creating commonsense knowledgebase

3.2 Semantic Role Identification

Semantic role is the relationship that the syntactic argument has with the verb. For each extracted sentences, the semantic roles should be identified. Different SRL (Semantic Role Labeling) tools like ASSERT (Automatic Statistical SEmantic Role Tagger) which requires Linux can be used for this purpose [8]. Consider an example

''The dog barked at a cat in the park last night".

There are mainly four semantic roles in a particular sentence. By using the ASSERT, it is possible to get these four semantic roles based upon the subject, object, verb, locative information and temporal information. For the previous example, ASSERT will give the result as the semantic roles in the sentence as follows:-

[ARG0 The dog] [Verb bark] [ARG1 a cat] [ARGM-LOC in the park][ARGM-TMP last night]

Semantic role labeling techniques automatically identify the different semantic roles of a sentence [2] .Even though the results of this SRL tool may not give the accurate results. The main reason of this is the different writing styles in the web pages. In order to increase the accuracy, verification strategy for the semantic roles should be done. For each crawled sentence, the semantic roles of it are kept in a database as a knowledge item. For a sentence with multiple verbs, the associated semantic roles for different verb are regarded as distinct knowledge items [5].

3.3 Semantic Role Verification

The semantic roles retrieved from the SRL may contain wrong semantic roles. In order to avoid this situation, semantic role substitution can be used. Semantic role substitution strategy mainly focuses on four semantic roles such as ARG0, ARG1, ARGM-LOC and ARGM-TMP where ARG0 represents the subject, ARG1 for object, ARGM-LOC for locative information and ARGM-TMP for temporal information. For the verification process, some fictitious sentences will be created [1]. And then evaluate each semantic role by substituting a specific semantic role in the given sentence. And then parse and compare the newly composed sentence with the original sentence. If both are equal, then that particular role will be taken for further processing. This process will continue until all the roles of each sentence are verified. By doing this, it is possible to increase the accuracy to above 90%.And all the roles which were verified should be stored in a database. In this stage, each semantic roles of the sentence are verified using substitution strategy. By analyzing the database, it is possible to see the verbs like "locate" and "find" give the highest instances of ARGM-LOC and the verbs like "see" and "get" gives the highest number of instances of ARGM-TMP. Consider those four verbs for the substitution strategy and then creating different substitution sentences with these verbs.

Then substitute the different roles retrieved from the ASSERT in these newly created sentences and repeat the semantic role identification process. After retrieving the semantic roles, check whether the roles retrieved in these two phases are similar or not. If the roles are different, then it is possible to assume that that particular sentence not at all considered as commonsense so that can be discarded. By doing this, it is possible to verify the different semantic roles retrieved and in this phase it it is easy to identify the sentences which will lead to commonsense.

3.4 Knowledge Distillation

After verifying the semantic roles, the next stage is to filter out the valid commonsense knowledge from the data retrieved so far. In order to identify the commonsense knowledge, different filtering rules can be applied and thereby it is possible to remove the unwanted items. Even after completed the semantic role verification, there will be some unreasonable commonsense. In this stage, that unreasonable commonsense knowledge will be removed. Sometimes, there will be number of words in the part of ARG1, then it is possible to assume that, the corresponding sentence refers to specialist's knowledge or sometimes it may be a meaningless sentence. After that a human will be going to evaluate those results. Since human can possess commonsense knowledge, the human can distinguish reasonable commonsense and unreasonable commonsense from the last results.

Then the reasonable commonsense is stored in a database and is referred to as commonsense knowledge base. This commonsense knowledge can be used for a variety of real life applications. As an illustration, the student behavior can be easily identified. On the other hand ,by using this knowledge the behavior of a student in a University can be easily assumed and can be presented in report format and can transfer this to the higher authorities or to the parents. This report will be created automatically based on

the different performance of the student in academics and extracurricular activities.

4. CONCLUSION AND FUTURE WORK

The knowledgebase created by the above framework can be directly applied to the different artificial intelligent systems. The main methodology used here is based on the integration of semantic role labeling and lexico-syntactic analysis. The lexico-syntactic pattern matching and semantic role labeling technique will help to improve the efficiency and the results will be more accurate.

In this work, after extracting the content from web pages, the content is given to the semantic role labeling engine. This engine will perform the semantic role labeling. If this particular content is possible to summarize or if it is possible to extract only sentences not all keywords, then it will lead to better results. For this purpose, it is inevitable to develop a natural language processing tool or a grammatical tool. Then the content from the first phase can be given to this grammatical tool or to this natural language processing engine, and then we can remove almost all unwanted words and sentence at the early stages of development. Then the results will be more accurate.

5. Acknowledgement

This paper has grown out as part of the fulfillment of the Masters Degree coarse project. we extend our thanks to Dr. R.Elijah Blessing Vinoth, M.E,Ph.D**.**, Director and The Head of the Department, School of Computer Science and Engineering for his encouragement and guidance**.** We would like to thank the reviewers for their insightment comments.

6. REFERENCES

[1] Sheng-Hao Hung, Chia-Hung Lin, Jen-Shin Hong," Web mining for event-based commonsense Knowledge using lexico syntactic pattern matching and semantic role labeling", *Expert Systems with Applications*, Volume 37, Issue 1, Pages 341-347, January 2010

[2] Palmer, M., Kingsbury, P., & Gildea, D."The proposition bank: An annotated corpus of semantic roles", *Computational Linguistics*, 31(1), 71–106, 2005

[3] [online] http://web.media.mit.edu/~hugo/montylingua.html

[4] Lenat, B. D., & Guha, V. R. ,"The evolution of CycL, the Cyc representation Language", *ACM SIGART Bulletin*, 2(3), 84–87, (1991).

[5] Gildea, D., & Jurafsky, D.," Automatic labeling of semantic roles". *Computational Linguistics*, 28(3), 245–288, (2002).

[6] Lenat, B. D. "CYC: A large-scale investment in knowledge infrastructure", *Communications of the ACM*, 38(11), 33–38., (1995).

[7] Miller, A. G. "WordNet: A lexical database for English.", *Communications of the ACM*, 38(11), 39–41., (1995).

[8] Pradhan, S., Ward, W., Hacioglu, K., Martin, H. J. & Jurafsky, D. ,"Shallow semantic parsing using support vector machines*.", In Proceedings of the human language technology conference/North American chapter of the association for computational linguistic*s (pp. 233–240),2004.

[9] Richardson, D. S., Dolan, B. W., & Vanderwende," MindNet: Acquiring and structuring semantic information from text", *In Proceedings of the 17 international conference on computational linguistic*s (pp. 1098–1102), 1998.

[10] Aone, C., & Ramos-Santacruz, M., "REES: A large-scale relation and event extraction system*", In Proceedings of the sixth conference on applied natural language processing* (pp. 76–83). Seattle, Washington, 2000.

[11] Liu, H., & Singh, P.," MAKEBELIEVE: Using commonsense knowledge to generate stories", *In Proceedings of the 18th national conference on artificial intelligence*, AAAI (pp. 957–958). Edmonton, Alberta, Canada, 2002.

[12] Liu, H., Lieberman, H., & Selker, T.,"GOOSE: A goal-oriented search engine with commonsense", *In Proceedings of the 2002 international conference on adaptive hypermedia and adaptive web based system*, Malaga, Spain, 2002.

[13] Liu, H., & Singh, P.,"ConceptNet – A practical commonsense reasoning toolkit", *BT Technology Journal*, 22(4), 211–226, 2004

[14] Liu H and Singh P, "Commonsense reasoning in and over natural language", *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES-2004).*

[15] Sheng-Hao Hung, Pai-Hsun Chen, Jen-Shin Hong, and Samuel Cruz-Lara," Context-based image retrieval: a case study in background image access for multimedia presentations", "IADIS International Conference WWW/Internet 2007.

[16] Hearst, A. M. (1992). "Automatic acquisition of hyponyms from large text corpora", *In Proceedings of the 14th conference on computational linguistics* (pp. 539–545).

[17] Nakamura, J., & Nagao, M. (1988). ,"Extraction of semantic information from an ordinary English dictionary and its evaluation", *In Proceedings of the 12th international conference on computational linguistic*s (pp. 459–464). Budapest, Hungry.

[18] Ponzetto, P. S., & Strube, M. "Semantic role labeling for conference resolution". *In Companion volume of the proceedings of the 11th meeting of the European chapter of the association for computational linguistic*s (pp. 143–146), 2006

[19] Palmer, M., Kingsbury, P., & Gildea, D., "The proposition bank: An annotated corpus of semantic roles." *Computational Linguistics*, 31(1), 71–106, (2005)

[20] Jensen, K., & Binot, J. "Disambiguating prepositional phrase attachments by using online dictionary definitions", *Computational Linguistics*, 13(3/4), 251–260, 1987