Next generation web - intelligent search, question answering, summarization and more

Emdad Khan

Abstract – Wouldn't it be nice to say or type "show me all the pictures from last Saturday party" on a browser and get all the requested pictures from Facebook? Or do specific transaction like "I would like to buy the following book - Artificial Intelligence by Stuart Russel, 3rd edition; please use my credit card on file and ship it to my home address (assuming that the user is already logged on to the specific website)" and receive the requested book on time? Or ask a question and get a specific answer? Or get summary of an article"? Or get a much better prediction from a BI (Business Intelligence) or Analytics software?

In fact, based on good research, we see a clear trend that the **future Internet** will be something that can provide **very specific, more precise and direct information (like the examples mentioned above) in a very easy way so that anyone including an illiterate person can access and use it at ease.** We call this **Intelligent Internet (IINT).**

Existing search engines usually provide thousands to millions of search results for any typical search. It is not easy even for advanced users to find the desired results from such a large set. One cannot get a specific answer or a set of answers to a question typed in a search engine. There is no automated way to get a good summary of a document or get a good inference from a document. Similarly, there is no way to get some specific desired information like "basic information of last 3 flights I took on United Airlines".

However, as mentioned, these are the key features that users would expect from next generation internet. Moreover, users would like to use such features in a natural way - like using a natural language sentence (by typing or preferably, by saying it; and for many cases using sentences that may not be grammatically correct). This is obviously a very complex task (and hence not solved yet). We would need multiple approaches, algorithms and technologies to achieve these. For example, Natural Language understanding (NLU), Big Data and Intelligent Agent (IA) are the 3 key areas we need to focus on. A Semantic Engine is the core engine that is needed for all these 3 major areas. In this paper we describe how a Semantic Engine using Brain-Like Approach (SEBLA) can address all key complexities of the next generation Internet and effectively provide all the key desired features mentioned above. We focus on Intelligent Search, Question Answering (Q & A System) and Summarization.

Dr. Emdad Khan is with the Dept. of Computer Science, Maharishi University of Management, 1000 N Fourth St, Fairfield, IA 52557, USA. He is also with InternetSpeech, Inc, San Jose, CA, USA (Phone: 408-532-9630, fax: 408-274-8151, email: emdad@internetspeech.com).

We also show how NLU, IA and Big Data can be integrated with existing client-server based web application architecture using the design patterns (e.g. Model-View-Controller) and software frameworks (e.g. Java or Ruby-on-Rails).

Keywords ---- Big Data; Natural Language Understanding (NLU); Semantics; Artificial Intelligence; Intelligent Internet; Intelligent Search; Question & Answer System; Summarization, Knowledge Extraction; Intelligent Agent; Machine Learning; Predictive Analysis; Business Intelligence; Information Technology.

I. INTRODUCTION

IN **this Information Age,** "information is money" like "time is money". The largest source of information is the Internet. Hence, it is important that everyone can easily and economically access the Internet, and effectively use all the key benefits of the Information Age.

As we know Internet has changed the world in a significant way. It is needless to mention the importance of the Internet for education; employment; economic, social, cultural & other developments; and more. Internet has become an essential part of life. It is a key driver for almost everything, including the basic necessities – food, water, shelter, health and the like. *We have seen the progression of the Internet from portal (Yahoo) to search (Google), to e-Commerce (e-Bay, Amazon) to social networks (Facebook, Twitter).* **What is NEXT?**

Well, we see a clear trend that the **future Internet** is going to be something that can provide **very specific, more precise and direct information in a very easy way so that anyone including an illiterate person can access and use it at ease.** This has TWO broad parts:

1. Natural User Interface (especially, easily imputing information, e.g. by naturally talking).

2. Retrieving more precise information – especially like much smaller set of search results, an answer or a small set of answers to a question, receiving specific information for a request, completing a specific transaction, getting a summary or drawing some inference from a document.

We call such a next generation Web or Internet as **Intelligent Internet, IINT** ([1], [2]).

Today, if we type something like

"Auto body shops in San Francisco bay area", we will get over million results - e.g. Google gave the following

"About 1,200,000 results (0,45 seconds) ".

Even though search engines are great (and Google Search is probably the best and most popular), and provide good ranking and hence display good results on the first few pages, such a large set of results is too many. An expert user can quickly decide which one to select from first few pages; and in many cases we get some good desired results. In many cases we need to rephrase the search words to get desired results. And in many cases we fail to get useful information. No one, of course, goes over the millions of hits as that is almost impossible - so what is the point of displaying such a large set? For non-expert users, such results are much less useful. And for illiterate or semi-literate people such a large set has far less value.

On the other hand, if we type a question like

"How many students graduated from Stanford University in Computer Science in 2012?" in any search engine, the results will be things like "Stanford university Palo Alto; Stanford university campus; computer science at Stanford etc etc". The real answer will not be there as the question was not understood to begin with let alone figuring out the answer.

And for other features like information on specific request

e.g. "**Show me all the pictures from last Saturday party",** there will be no useful results. The same is true if we ask for a summary or inference from a document.

The key reason we do not get much smaller search results or answer to a question and the like is that the problem is very complex. In fact, there are multiple **complex problems, namely, understanding human language, discovering knowledge & intelligence from very large data [so called Big data ([3] - [7])], formulating answer, and acting properly to user's request.**

We would need multiple approaches, algorithms and technologies to solve these complex problems. For example, Natural Language understanding **(NLU),** Big Data and Intelligent Agent **(IA)** are the 3 key areas we need to focus on. A Semantic Engine is the core engine that is needed for all these 3 major areas. In this paper, we describe how a **Semantic Engine using Brain-Like Approach (SEBLA)** ([2], [9]) can address all key complexities of the next generation Internet (IINT) and effectively provide all the key desired features mentioned above.

NLU is needed for multiple purposes including a simple user interface (**UI**), understanding user's request, and retrieving & formulating desired answers. Big Data is essential to deal with key Big Data related issues, most importantly discovering knowledge and intelligence from data (NLU is essential for this as well, as data on the web is dominated by unstructured data; however, NLU is also needed to deal with structured data and integrating it with unstructured data [1]). Intelligent Agent (IA) is needed to properly understand all tasks and execute them.

From implementation standpoint, it is important to note that NLU and IA can be implemented in the server, client or both in the client-server architecture of the web. In general, for complex cases, these would be needed in both server and client. The Big Data processing capability would be mainly in the server; however, for some complex cases, client may also play a good role.

Many scientists, engineers, researchers and others have been working on Advanced Search, Question & Answer (Q&A) system, NLU, Semantic Engine, Big Data and Intelligent Agent. They have been using various algorithms, methods and technologies and have made outstanding contributions.

In this paper we have addressed these problems in two critical ways:

1. Determining the "core" problem to all these and developing a "core" engine.

2. Developing an approach (along with architecture, algorithms and the "core" engine) to solve these complex problems using Brain-Like and Brain-Inspired algorithms (as human can easily deal with such problems).

Our "core" semantic engine SEBLA uses the semantics of words to derive the semantics of a sentence, and the semantics of sentences to derive semantics of a paragraph. It uses "Deep Learning", "Deep Semantics" along with an integrated approach to address all associated complex problems.

Section II describes key issues related to Intelligent Internet (IINT - [1], [2]). It includes some details of a Question Answering system. Section III describes our approach using SEBLA to address the key issues with IINT. Section IV describes how our approach is used to provide Intelligent Search.

Section V describes how to integrate NLU, IA and Big Data with Web Application Architecture. Section VI describes how

our solution can be used in various other applications, especially for Summarization, and Section VII provides Conclusions & Future works.

II. ISSUES RELATED TO INTELLIGENT INTERNET, IINT

Intelligent Internet, IINT is described in details in [2]. A website in IINT would need to have one or more Intelligent Agent (IA). There can also be Super Intelligent Agent (SIA) to do more complex higher level tasks and collaborating with the lower level IAs on different websites. Today's websites do not have any IAs although some websites do have some of the functions of IA – e.g. using web services. The environment of an IA is basically the HTML/XML content of the site, content of other sites if those sites do not have any IA, IAs of other sites, SIAs and the users. Thus, IINT is a system of IAs and SIAs, usually, working in co-operation (unless we are talking about two competitive sites). It is important to note that each website would need to have an IA to make the website much more effective – *a major paradigm shift in website / web application design and implementation*.

The major functions of IA are to interact with the user, understands user's request, make appropriate decisions, ensure all critical tasks / sub tasks are completed, and act to serve the user. Such actions involve collaborating with IAs and SIs of other websites, accessing appropriate databases (SQL, NoSQL and the like), accessing content from other websites, extracting relevant content, formulating the results and presenting it nicely to the user. Clearly, for all these, a Semantic Engine (e.g. SEBLA) is essential.

An example will make it more clear. Let's consider a Question and Answer (Q&A) system. If we say or type our example question mentioned above i.e.

"How many students graduated from Stanford University in Computer Science in 2012?", *then an IA will be doing the following specific tasks:*

1. Convert speech to text if the sentence was spoken.

2. Derive meaning (semantics) of the sentence using the semantics of the words in the sentence.

3. Determine all the actions needed based on the meaning of the sentence i.e.

a. It is a question.

b. It needs access to Stanford University website.

c. It needs "world knowledge" (in this case it can be a Name Entity Recognition, especially if the word "university" was not there or just a table showing Stanford University website address).

d. Go to Stanford University web address www.stanford.edu.

e. Do on-site search "graduated students computer science".

f. Use the meaning of all search results and click the appropriate link.

g. Determine that an account is needed with login-password; ask the user accordingly to provide login information.

h. User will then provide account information (to make it simple, let us say there is no Capcha or equivalent verification question).

i. Once logged on, IA will try to find the database and then appropriate table(s) and fields.

j. IA then will retrieve the requested information and check it with the input question. It will retrieve more information if the answer is not well related to the question.

4. IA will then formulate a short user friendly answer and present it to the user.

The tasks of IA for Intelligent Search, request for a Specific Information (like getting all pictures of the last weekend party), Summary of an article and the like will be similar at the top level (mainly getting the semantics and using it properly for the detailed tasks) but will vary at the implementation level. For some tasks, SIAs will be involved to communicate and collaborate with other SIAs / IAs at other websites.

Thus, it is clear that the SIAs and IAs would need five types of high level functions:

1. Understand user's request using semantics.

2. Determine actions.

3. Communicate with other sites.

4. Retrieve data and address Big Data issues (especially, extracting knowledge and intelligence from data; again, using semantics / Semantic Engine as the major technology) as appropriate.

5. Formulate results and present it to the user.

Unstructured data dominates the data world. It is estimated that over 80% data in computers and Internet are unstructured

[3]. Unstructured data can be broadly classified into two groups:

(a) Text data and
(b) Non-text data (including sound, image video).

Computers are very good in processing structured data (e.g. data in a database). This is mainly because computers are still mathematical devices, especially, fast number crunchers. When it comes to unstructured data, we are dealing with the meaning or semantics and associated context; and humans are very good at that.

In the textual case there is a key problem of context. The classic example often given is the difference between the statements that

"John rides in a mustang" and "John rides on a

mustang" [7].

A human analyst will see a great deal of difference between these two sentences. Our experience adds enormously to our understanding of both. We know, for a start, that the first statement refers to a car, the second to a horse. But we will also understand that in the first statement John is a man, and he is probably in the United States, because Ford Mustangs are not sold in large numbers outside the US.

In the second statement, we may consider that the event might have occurred in the US as the descriptive term is generally associated with that country, and a long time ago, as there are not many wild horses left in the US. It might even occur to us reading one of the two sentences that, because the O and I keys are beside one another on a standard keyboard, there could have been a typographical error and the other sentence may be the correct one.

The human brain picks up all of this data almost instantaneously – our understanding is implicit. Computers cannot deal with implicit information and have to be told how to understand it. Consequently they deal with this 'tacit' information very badly, if at all.

This gets further complicated as the writing style, sentence structure and vocabulary used in formal documents are very different to those used in e-mails, which are in turn different to those used in text messaging. Humans can handle all these very well.

One classical approach used in a computer to handle text data is "keyword" or key phrase search. Although useful, this method is far from perfect. If the set of search terms is too narrow it can miss vital information, if it is too broad the resulting set of 'hits' can contain large numbers of totally irrelevant 'false positives'.

Modern search tools have improved things somewhat. Computerized thesauruses allow us to search for synonyms and homonyms without having to explicitly set out every possible variation. Other tools allow for 'stemming' - for example, in Lexis Nexis putting in the term 'run+' will cause the engine to search for 'run', 'runs', 'running', 'runner', and so on.

One key modern method is the use of some semantics using Predicate logic, Ontology and the like. However, one would need to define clearly all such semantics. Any small variation in the words or structure can cause the semantics to be different yielding wrong results or no results. Such approaches basically provide some "mechanical" semantics; thus limiting them to applications with small domain.

The problem becomes even more critical when we try to use non-text data – like audio, image, video. Here also, human brain handles such data very efficiently.

Thus, existing approaches have simplified the process somewhat, but they still have not solved the problem of computers' inability to deal with tacit and context-based information. At present, we can conclude that text analysis technology may be better at data reduction than actual data analysis. As already explained, human brain is very good in addressing these problems. The key point is that we would need to use the semantics and NLU capabilities in dealing with unstructured data (see Section III).

It is important to note that although humans can do text processing very well, they can do it only for relatively smaller size data. Human brain cannot handle very large data like big data. **However, using human brain's intelligent approach with the fast number crunching computers, we believe, we can effectively solve the big data problem - the theme of our approach.**

III. SEMANTIC ENGINE AS THE CORE ENGINE TO ADDRESS KEY PROBLEMS OF IINT

Clearly a Semantic Engine (e.g. SEBLA) is the key to address all the complexities related to IINT. Since unstructured data dominates with a wide margin, a brief description of SEBLA in the context of unstructured data is provided below (SEBLA also handles structured data ([1], [2]).

The key problems associated with unstructured data are related to the semantics of words, sentences and paragraphs. Human brain uses semantics and natural language understanding (NLU) to very efficiently use unstructured data.

Below, first we briefly describe a Semantic Engine ([2], [9]) using Brain-Like algorithms (SEBLA). Then we show how SEBLA can handle some Big Data applications, namely, Intelligent Search, Summarization and IINT itself.

3.1 **Semantic Engine - SEBLA**

While traditional approaches to NLU have been applied over the past 50 years and had some good successes mainly in a small domain, results show insignificant advancement, in general, and NLU remains a complex open problem. NLU complexity is mainly related to **semantics**: abstraction, representation, real meaning, and computational complexity. We argue that while existing approaches are great in solving some specific problems, they do not seem to address key Natural Language problems in a practical and natural way. In [9], we proposed a Semantic Engine using **Brain-Like approach (SEBLA)** that uses Brain-Like algorithms to solve the key NLU problem (i.e. the semantic problem) as well as its sub-problems.

The main theme of our approach in SEBLA is to use each word as object with all important features, most importantly the semantics. In our human natural language based communication, we understand the meaning of every word even when it is standalone i.e. without any context. Sometimes a word may have multiple meanings which get resolved with the context in a sentence. The next main theme is to use the semantics of each word to develop the meaning of a sentence as we do in our natural language understanding as human. Similarly, the semantics of sentences are used to derive the semantics or meaning of a paragraph. The 3rd main theme is to use natural semantics as opposed to existing "mechanical semantics" of Predicate logic or Ontology or the like.
A SEBLA based NLU system is able to:

1. Paraphrase an input text.
2. Translate the text into another language.
3. Answer questions about the content of the text.
4. Draw inferences from the text.

As an example, consider the following sentence:

"Maharani serves vegetarian food."

Semantics represented by existing methods, e.g. Predicate Logic, is

Serves(Maharani, Vegetarian Food) and

Restaurant(Maharani).

Now, if we ask
"Is vegetarian dishes served at Maharani?"
the system will not be able to answer correctly unless we also define a semantics for "Vegetarian Dish" or define that "food" is same as "dish" etc. This means, almost everything would need to be clearly defined (which is what is best described by "mechanical semantics"). But with SEBLA based NLU, the answer for the above question will be "Yes" without adding any special semantics for "Vegetarian Dish".

The "mechanical semantics" nature becomes more prominent when we use more complex predicates e.g. when we use universal and existential quantifies, and/or add constructs to represent time.

3.2 How SEBLA Works?

The Semantic Engine using Brain-Like Approach (SEBLA) has two phases [9]:

a. "training" phase and
b. "recall" phase.

The "recall" phase can also have the on-line training as a continuous learning process, mainly for refinement. Fig. 1 shows the training phase. An input sentence (or words) is provided to both SEBLA and Semantic Database (SD). SD is formed by taking a natural language sentence corpora and defining semantics by some Natural Language experts. The format of the semantics may vary. Since SEBLA derives semantics of a sentence using semantics of words and natural language grammar, ideally a relatively small SD is needed. In some cases - e.g. for small vocabulary applications, an SD may not be needed at all. So, SD is mainly used for refinement of SEBLA semantics. Since a natural language corpora may have trillions of sentences (e.g. Google Trillion Sentence database), manually creating Semantics for all sentences in such a large corpora is a daunting task. Hence, we recommend to use a small size SD to ensure that SEBLA has a good start and then use just SEBLA itself and (when needed) an SD for refinement. However, instead of manually creating a large SD, all sentences that SEBLA correctly derives semantics during its Recall phase over time, can be automatically added to the SD. This can be done by properly monitoring SEBLA's performance on a continuous basis (Fig. 2; more details in Section IV).

For any incorrect performance of SEBLA, it tries to re-train itself. Retraining includes fine tuning of the semantics of the words as appropriate. A large language corpora is still useful in three major ways: to refine and enhance the World Knowledge (WK), to refine and enhance **Function Words** of a word (Function Words of a word are the key feature words that describe the function of a word. E.g. for the word "ball", the function words are {ball, move, roll, round, play..} - [9]), and to help better use of the grammar of the language.

The Machine Learning (ML) box shown in Fig.1 and Fig. 2 can use almost any machine learning algorithm as appropriate.

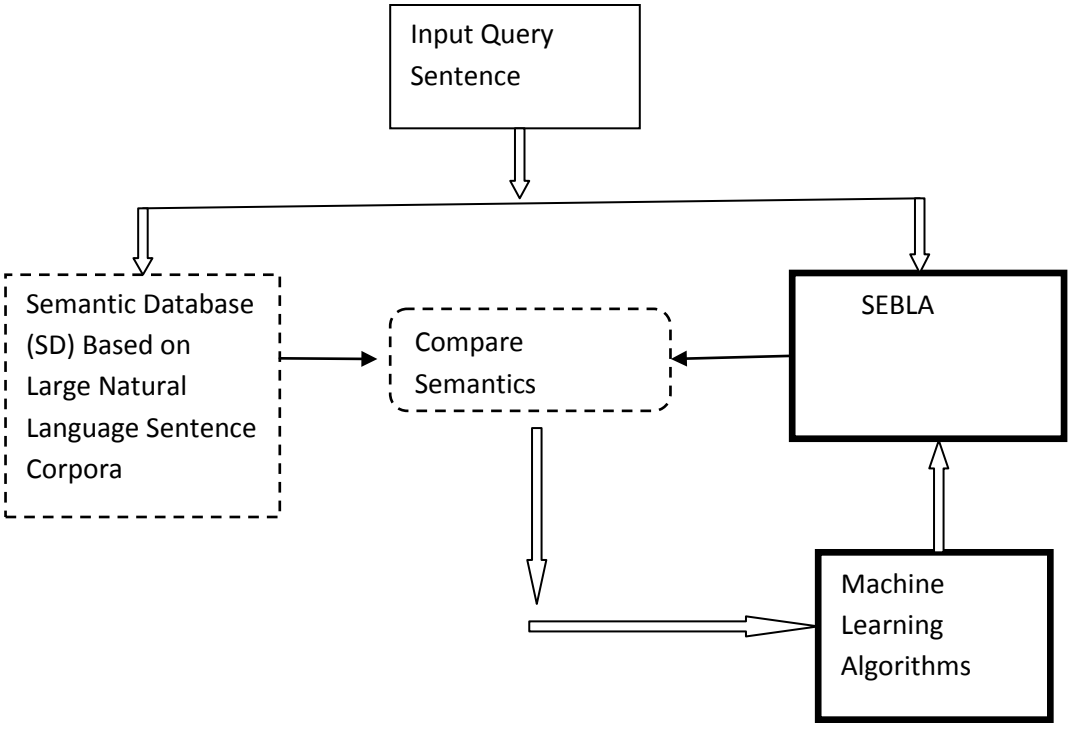

Fig. 1: Semantic Engine Using Brain-Like Algorithm – **Training phase** using large Natural Language Sentence Corpora and Machine Learning Algorithms.

At the word level, the key question we have addressed is how to represent the semantics for each word (see the example shown for the word "ball" above) and how to associate appropriate world knowledge with each word. By using the representation and semantic feature of each word, along with the world knowledge associated with each word [9], the meaning of a sentence is derived by applying the grammar of the language and appropriate rules to combine words. Key features of the words and appropriate rules to combine them are learned / refined using large text corpora and machine learning algorithms. The inference engine (Intelligent Agent - Fig. 2) determines the meaning / semantics of a sentence by using the word semantics and appropriate rules to combine the words in a sentence.

IV INTELLIGENT SEARCH USING SEBLA

The information retrieval process through existing search engines and Information Retrieval systems are mainly based on string match. Thus, the search process needs to deal with many string comparisons to find matches. And all matched data are extracted even though many data are not relevant and desired. Accordingly, such engines produce many (*often millions of*) results as described in Section I, e.g., if we type something like

"Auto body shops in San Francisco bay area",

we get over million results - e.g. Google gave the following:

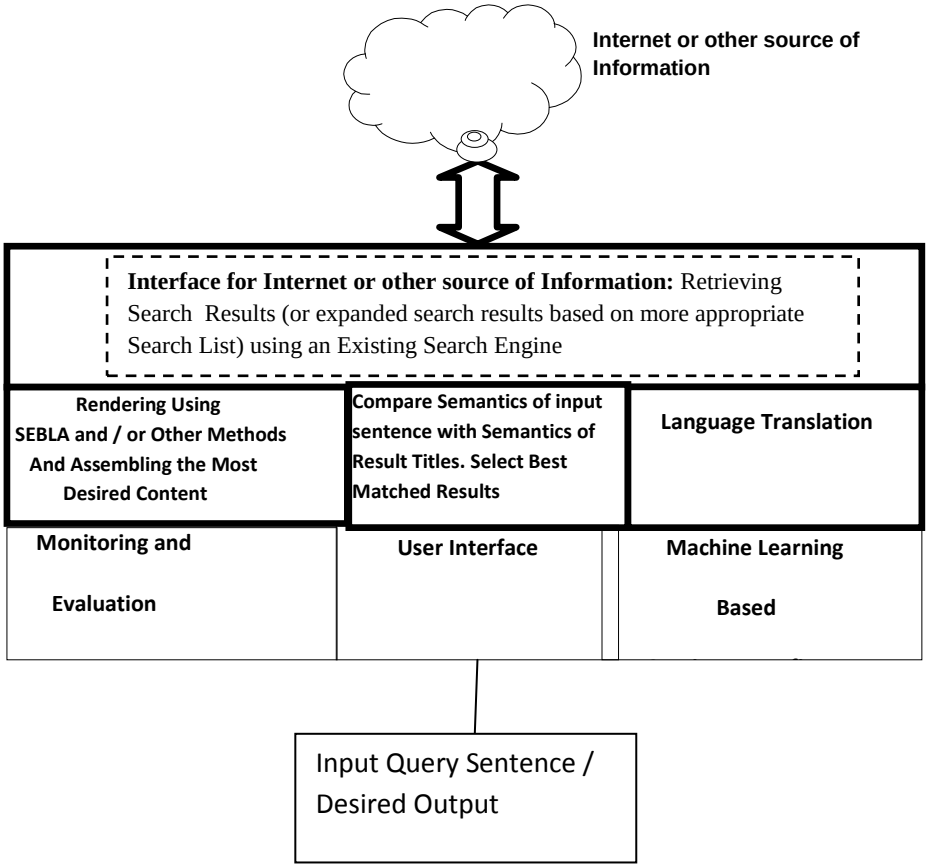

Fig. 2 Semantic Engine Using Brain-Like Algorithm – showing complete retrieval of desired information. The Intelligent Agent (IA) shown is for Intelligent Search using SEBLA, SEBLA_IS. This also represents the Recall Phase corresponding to Fig. 1.

" About 1,200,000 results (0,45 seconds) ".

Human knowledge and intelligence are needed to retrieve the desired information from such large search results. This requirement usually limits the usage of search engines to experienced and educated users. There are **four** key issues with the current approaches:

a. Search process needs to deal with very large data.
b. String search results contain many undesired and unrelated results.
c. String search results may not contain the desired results and user may need to do multiple searches by various search word combinations.
d. String search results may NOT contain the desired information even after trying major keyword combinations as a user may skip key words of similar meaning.

The semantic capability of SEBLA addresses these issues in multiple ways.

One simple way is to Retrieve all results using existing string match based search from an existing search engine. Then apply semantics to all the titles / headlines and compare the semantics of the titles / headlines with the semantics of the input sentence or search string, and then select the results based on high semantic matching.

The key steps using this approach are (**Fig.** 2):

1. Calculate the meaning / semantics of the query words and sentence.
2. Calculate the semantics of each title / indexed item of the search results, and then calculate semantic matching or overlap of the query with each target.
3. Then select the titles with high semantic matching and present to the user.

Let's consider the following search string:

"Low price Thai restaurants in Silicon Valley"

to Bing which produced **1.2M** results (Google produced 1.1M results). After applying SEBLA based search engine (as shown in Fig. 2), we got only **47 results** (Fig. 3).

This sentence does not need to access a database (i.e. no account information is needed) but it uses some words the meaning of which are subjective (e.g. low price, Thai). Besides, it needs a good world knowledge (WK) as the intention of the search is clear to us but not to the computer i.e. the intention is to get some names of good Thai restaurants in Silicon Valley. So, this represents a good sentence to test the semantics and quality of the final results from existing search engines.

How did SEBLA based Intelligent Search (**SEBLA_IS**) got 47 results while string search retrieved over million results? As we know, there are not 1.2 million Thai restaurants even in the whole world!

As mentioned, string search retrieves almost everything matched but qualified with number of occurrences of words, inverse document frequencies (IDF), some probabilities based on Natural Language Corpora and some standard semantics that existing search engines are using. Hence, it gets such a large hits. E.g. it lists all the news papers that mentioned the name(s) of a Thai restaurant - may be as a review or as an article etc. But that was not the intention of the input search string.

SEBLA_IS used semantics and the world knowledge (WK) to filter out all results that are not relevant. It is evident from Fig. 3 results. *The results shown in Fig. 3 are based on just word semantics (sort of like 1-gram semantics). Thus, quite a few unwanted results are still there (e.g. 4th results from the top [left column] has Thai Ice Tea). For **2-consecutive** word semantics (sort of bigram semantics), the total number of results came **down to 26**). With higher number of consecutive word semantics, we would get more accurate results.* If SEBLA_IS used its **own semantic search** i.e. compare the semantics of possible hits with the semantics of input query before retrieving it, it would get even better results. SEBLA_IS also used "Deep Semantics" and "Deep Learning" more like in a natural way i.e. not using the "Mechanical Semantics" commonly used in existing methods mentioned before. Note that some results are repeated in Bing and we kept it as is; but such repeated results can easily be discarded.

Another way to do Intelligent Search is to understand the meaning of each word and sentence in the query sentence and then perform the following:

1. Generate all equivalent sets of query strings of the input sentence (thus generating lot more appropriate search results that are related to the input words and sentences).

2. Extract the most appropriate and related results from the extended search results as described in Fig. 2.

There are a few other similar methods as well.

Fig. 2 also shows a few other blocks, namely, "rendering", monitoring, language translation, machine learning and continuous improvement. Among these, the "**Rendering**" ([10], [11], [12]) box (the middle left box) in Fig. 2 deserves some explanation which is provided below.

If any search results shown in Fig. 3 is clicked, the SEBLA_IS will show ONLY the **desired content** on the new page instead of showing unrelated content of the new page. As we know, the Internet was designed with visual access in a relatively large display screen (like a 8.5 inch x 11 inch page) in mind. Thus, all the content are laid out on any website and webpage in a manner that attract our eyes in a large screen. Retrieving the desired content (which is much smaller in size than the total content on a webpage or website) from a typical webpage / website and displaying that (or playing in audio) into a much smaller screen (like in a cell phone or PDA) is a very challenging task. This process of retrieving and converting most desired content from a large source of content into a much smaller but desired content is called **"rendering".** Clearly, rendering is mainly related to Internet Browsing on a small device. An Intelligent Search uses rendering to provide very specific desired content. Rendering includes form rendering, retrieving appropriate data when a form is submitted, and retrieving multi-media data.

V INTEGRATING NLU, IA AND BIG DATA WITH WEB APPLICATION ARCHITECTURE

It is important that we can seamlessly integrate NLU, IA and Big Data with existing client-server based web application architecture using the design patterns (e.g. Model-View-Controller, MVC) and software frameworks (e.g. Java or Ruby-on-Rails). Existing architecture with MVC uses Controller to accept user's input and performs specific tasks (e.g. completing an e-Commerce transaction) using the Model and present the results using the View. The specific way of performing these tasks depend on the software framework used.

To understand user's input (query, request etc) using natural language, we need to integrate a NLU engine in the Controller (in fact, enhanced Controller). NLU (using Semantic Engine, e.g. SEBLA) will create appropriate

InternetSpeech Intelligent Search Results are:

BEST RESTAURANTS FOR DATES: SILICON VALLEY ... (Not sure about price because menu was taken ... Downstairs ambiance was nice with the low hung light â€¦

... silicon valley restaurant reviews | milpitas ... Barber Ct, Milpitas, CA. 408.526.9888. Pho Tam Thai/Vietnamese. $$. The attractive room belies the low prices.

It was already a famous Silicon Valley restaurant and watering ... low prices , no beer taps (but ... (like the 1980 invasion and conquest of the Bay Area by Thai ...

They have all sorts of Vietnamese fair for really low prices. ... small #9 & a thai ice tea is what i always ... Recent topics in "Silicon Valley" Topic Author Replies

Schmap San Jose and Silicon Valley Restaurants ... but the prices are low and the lunch buffet is a ... Restaurants - Thai San Jose and Silicon Valley - Restaurants

home | restaurants | silicon valley restaurant reviews ... Thai/Vietnamese. $$. The attractive room belies the low prices.

Amber India pretty much owns the Indian fine dining category in Silicon Valley. ... Best Thai Restaurant. ... Delicious, creamy, low fat and just as refreshing ...

San Jose, California dining ... Steak Restaurants; Thai Restaurants; Vegetarian Restaurants; ... Voted Best Steak restaurant by Silicon Valley residents, ...

... silicon valley restaurant reviews | milpitas ... Barber Ct, Milpitas, CA. 408.526.9888. Pho Tam Thai/Vietnamese. $$. The attractive room belies the low prices.

It was already a famous Silicon Valley restaurant and watering ... low prices , no beer taps (but ... (like the 1980 invasion and conquest of the Bay Area by Thai ...

They have all sorts of Vietnamese fair for really low prices. ... small #9 & a thai ice tea is what i always ... Recent topics in "Silicon Valley" Topic Author Replies

Schmap San Jose and Silicon Valley Restaurants ... but the prices are low and the lunch buffet is a ... Restaurants - Thai San Jose and Silicon Valley - Restaurants

home | restaurants | silicon valley restaurant reviews ... Thai/Vietnamese. $$. The attractive room belies the low prices.

Amber India pretty much owns the Indian fine dining category in Silicon Valley. ... Best Thai Restaurant. ... Delicious, creamy, low fat and just as refreshing ...

San Jose, California dining ... Steak Restaurants; Thai Restaurants; Vegetarian Restaurants; ... Voted Best Steak restaurant by Silicon Valley residents, ...

BEST RESTAURANTS FOR DATES: SILICON VALLEY ... (Not sure about price because menu was taken ... Downstairs ambiance was nice with the low hung light â€¦

... silicon valley restaurant reviews | milpitas ... Barber Ct, Milpitas, CA. 408.526.9888. Pho Tam Thai/Vietnamese. $$. The attractive room belies the low prices.

They have all sorts of Vietnamese fair for really low prices. ... small #9 & a thai ice tea is what i always ... Recent topics in "Silicon Valley" Topic Author Replies

Schmap San Jose and Silicon Valley Restaurants ... but the prices are low and the lunch buffet is a ... Restaurants - Thai San Jose and Silicon Valley - Restaurants

Amber India pretty much owns the Indian fine dining category in Silicon Valley. ... Best Thai Restaurant. ... Delicious, creamy, low fat and just as refreshing ...

[Some Results Taken Out to Fit in One Page]

Total Count of Results: 47

Fig. 3 Search Results for "*Low price Thai Restaurants in Silicon Valley*" With **Intelligent Search** (SEBLA_IS) Using **SEBLA** (Semantic Engine Using Brain-Like Approach) - Courtesy InternetSpeech Corporation (www.internetspeech.com).

meaning of the input. Then we need to integrate an IA (Intelligent Agent) to create appropriate higher level tasks to be performed (e.g. specific tasks of a Question Answering System described in Section II). These specific tasks may still be at a higher level than the Controller (of MVC) can handle. So, IA would need to break them at an appropriate lower level so that the Controller can perform them easily. The results (and additional processes) of the specific tasks would need to be managed by IA and present to the View block. Such tasks may involve some Big Data processing using some Big Data engine. Thus, the MVC would need to be enhanced to properly integrate NLU, IA and Big Data. But it is doable and specific implementation would need to enhance existing software framework.

VI OTHER SAMPLE APPLICATIONS

Our SEBLA and NLU based approach can be used in various other applications including Intelligent Information Retrieval, Q & A System, Summarization, Machine Translation and Business Intelligence. We describe Intelligent Summarization ([1], [9]) and how to implement IINT below.

6.1 Intelligent Summarization

Summarization process needs to consider the semantics of multiple sentences and multiple paragraphs. Semantics for multiple sentences and paragraphs can be calculated using SEBLA as it was done for calculating semantics of a sentence. However, some modifications are need for the following reasons:

1. Within a sentence, words are used in a constrained way using grammar. But between sentences there is no such grammar.
2. Usually, a group of sentences carry a theme within a context and there are ***relations*** between sentences.

Thus, to calculate the semantics between sentences, we will use word semantics as before BUT with some modifications. This is also true for a single long sentence segmented by "comma", "semicolon", "but", "as" and the like. We also need to take account **for "discourse" i.e. coherence or co-reference to words in previous sentences**. There are some good existing solutions mainly for a small domain problem. But, in general Computational Discourse (CD) in natural language is an unsolved problem. However, with our SEBLA based scheme, the CD problem can be solved to a good extent for large domains.

In calculating semantics in a long sentence, the previous, next and other words can further influence / refine the semantics. For convenience, we have included this aspect in calculating semantics of multiple sentences. Fig. 4 shows key steps in calculating semantics of multiple sentences. It also shows how summarization process is done.

Now let's consider semantics of the following paragraph:

"The Intelligent Internet (IINT) will take the Internet to a new level (S1). It will allow existing as well as significant number of new users to enjoy the existing and various new benefits of the Internet (S2). IINT will affect their lives in a positive way with Economic, Social, Cultural and other developments globally (S3)". …….…….....(1)

The core semantics of each sentences are as follows:

S1 -> {take internet} …….(2a)
S2 -> {allow enjoy} ……(2b)
S3 -> {affect lives} ……(2c)

(Note, we have used the Action Words for the core semantics [9]). Using the Function words in (2), we see relation between "internet" & "enjoy" (2a and 2b) and "internet" & "lives" (2a and 2c) and "enjoy" & "lives" (2b and 2c). Thus, the core semantics of these sentences can be represented as

{internet, enjoy, lives} …….....(3)

Note that we are looking for relationship with all 3 sentences. While "take" goes with "allow", it does not go with "affect". So, to calculate semantics using minimal relationship (so that we can get shortest possible summary), first we take minimal action words from all sentences.
Also, we are looking for repeat occurrences so that we can drop them. Thus, S1 should be dropped from this standpoint as S1 matches with S2 and S3 using the word "internet"

Now, we can derive the summary by using these core semantic words, matching with input sentences and compressing them, yielding

[Theme for summarization:

a. Sentences with no semantic overlap with any other sentences, should be kept unless its semantic value is relatively low.
b. Remove sentences having similar meaning (keep just one such sentence)
c. Delete sentences with relatively low semantics.

To achieve this we need to use the semantics of sentences using minimal action word overlaps etc as shown in this example]

"The Intelligent Internet (IINT) will allow existing as well as significant number of new users to enjoy the existing and various new benefits of the Internet. IINT will affect their lives in a positive way with Economic, Social, Cultural and other developments globally" …….....… (4)

The first sentence is dropped as the action word "take" did not match with any similar words *(see note above; also although "take Internet" is there, in general, function*

words of "take" does not have overlap with the function words of "internet") and hence not present in (3). However, the first sentence can be added if evaluation (see below) provides a low score to the "summary"

used to further improve the summarization. Besides, simple reduction can also be used e.g. "as well as" in S1 can be replaced with "and".

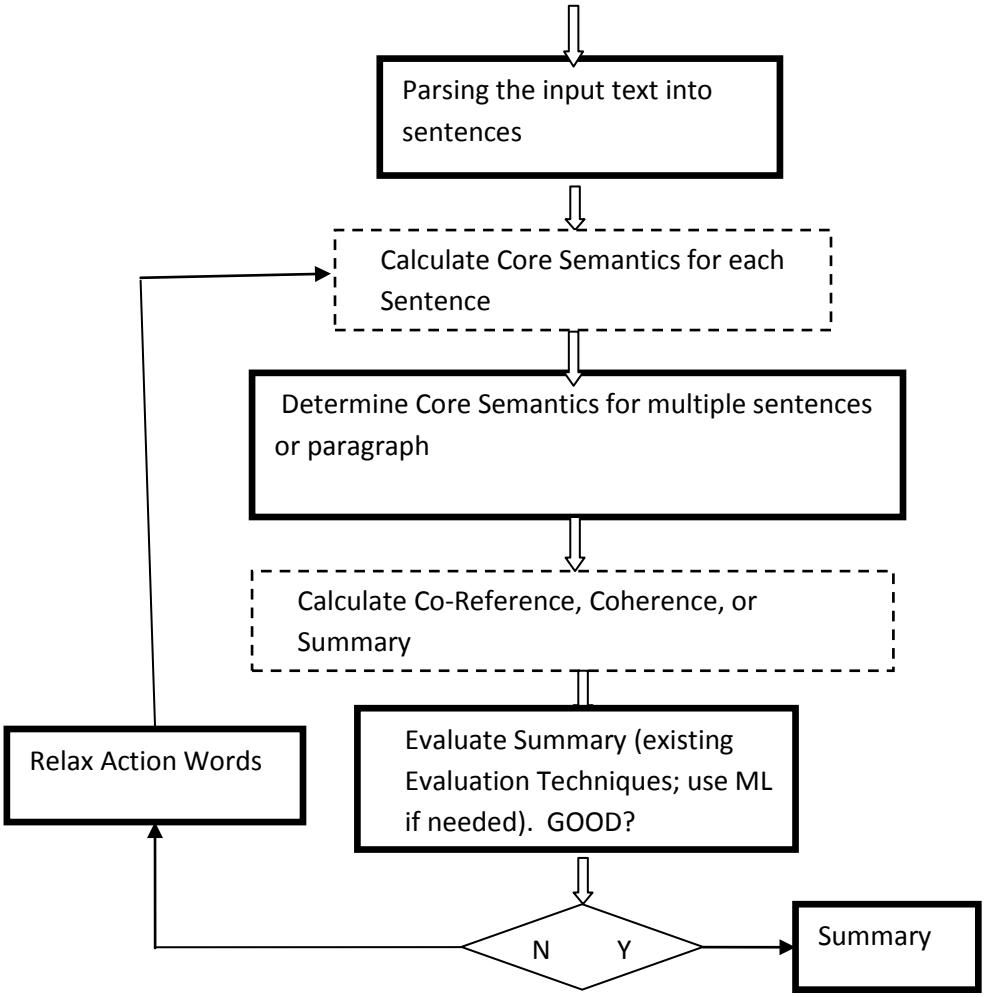

Fig. 4: Calculating Semantics for Multiple Sentences. If Summarization is calculated, the feedback loop is used for further refinement when needed.

(i.e. for cases where dropping a sentence (s) may lower the score too much). In general, a sentence not having any action does not belong to the "summary".

This may not be the best summary. In general, summarization is an iterative process – take minimum words for action(s) and associated object(s) and find the core semantics. Then calculate the summary. Then evaluate the summary using some evaluation techniques (including existing standard evaluation techniques e.g. ROGUE, Pyramid Method). If the evaluation score is low, relax the "action words" i.e. take more action words and associated objects and repeat the process. E.g. if we take both actions words "allow" and "enjoy" in S2, then "allow" will match with "take" in S1. Then we will have S1 in the summary etc. Additionally, machine learning (ML) can be

The same process can be applied to more sentences i.e. multiple paragraphs as paragraphs are joining of multiple sentences. If there is no good semantics between two paragraphs, then those two paragraphs are talking about different things and cannot be summarized i.e. we need to keep them as is or use higher level semantics (e.g. in drawing inferences).

SEBLA can help all key tasks in Summarization **including general summarization, question specific summarization, and creating abstracts.**

6.2 Implementing Intelligent Internet (IINT)

As mentioned before, the **future Internet** will be something that can provide very specific, more precise and direct

information in a very easy way so that anyone including an illiterate person can access and use it at ease. **The 3 key applications** described so far in reasonable details, namely, Intelligent Search, Q & A System and Summarization, are the key drivers for IINT along with Intelligent Agent and Big Data.

With IINT, users will be able to get much smaller set of search results (usually under 50), answers to questions, specific information on special request (like all pictures from last Saturday party), summary of a document, inference (extracting knowledge and intelligence) of a document(s) and more, using natural language based interaction - either spoken or typed; thus enabling a natural dialogue with the Internet and computers and getting most desired results.

VII CONCLUSIONS AND FUTURE WORKS

We have shown how the future Internet would look like based on the general trend, especially, considering the progression of the Internet from the beginning till today - we have seen the progression of the Internet from portal (Yahoo) to search (Google), to e-Commerce (e-Bay, Amazon) to social networks (Facebook, Twitter). We see a clear trend that the **future Internet** will be something that can provide **very specific, more precise and direct information in a very easy way so that anyone including an illiterate person can access and use it at ease.** We call this **Intelligent Internet (IINT).** IINT will support Intelligent Search, Q & A, Summarization, Knowledge & Intelligence Extraction and more.

The key drivers of IINT are Semantic Engine & Natural Language Processing, Big Data and Intelligent Agent (IA) with a Semantic Engine being the core engine for all these drivers.

It is important to note that NLU, IA and Big Data can be integrated with existing client-server based web application architecture using the design patterns (e.g. Model-View-Controller, MVC) and software frameworks (e.g. Java or Ruby-on-Rails).

The Impact of IINT is huge and multifold. It will enable all population group (rich, poor, literate, illiterate, blind, elderly and others) to more effectively access and use the information. All devices including mobile, tablet, laptop, desktop will be able to use IINT with visual and audio. IINT will take Internet to a new level and will allow existing users as well as significant number of new users to enjoy the new benefits of the Internet, and affect their lives in a positive way.

IINT will be a key driver for global development – economic, social, cultural and more via Education (including technical), Innovation and Entrepreneurship. It will also result in increased Global Peace.

Future works include Q & A System for a very large domain, more natural summarization, and more capable knowledge & intelligence extraction.

REFERENCES

[1] E. Khan, "Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach**",** INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) January 2014.

[2] E. Khan, " Intelligent Internet: Natural Language and Question & Answer based Interaction", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) Oct. 2013.

[3] C. Eaton et al, "Understanding Big Data: Analytics for enterprise class Hadoop and Streaming Data", http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF

[4] T. White, " *Hadoop: The Definitive Guide*", O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.

[5] J. Dean et al,

"MapReduce: Simplified Data Processing on Large Clusters", OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[6] Wikipedia – "Big Data" - http://en.wikipedia.org/wiki/Big_data

[7] P. Ryan et al, "The Problem of Analyzing Unstructured Data", Grant Thoronton, 2009, http://www.grantthornton.ie/db/Attachments/Publications/Forensic_&_inve/Grant%20Thornton%20-The%20problem%20of%20analysing%20unstructured%20data.pdf

[8] E. Khan, "Natural Language based Human Computer Interaction: a Necessity for Mobile Devices", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) Dec. 2012.

[9] Khan, E., (2011): Natural Language Understanding Using Brain-Like Approach: Word Objects and Word Semantics Based Approaches help Sentence Level. A Patent Filed in US in 2011.

[10] E. Khan, "Internet for Everyone – Reshaping the Global Economy by Bridging the Digital Divide", Book- ISBN 978-1-4620-4251-7(Soft Cover), 978-1-4620-4250-0 (Hard Cover), Aug 2011.

[11] E. Khan & E. Aleisa, "e-Services using any Phone & User's Voice: Bridging Digital Divide & help Global Development", IEEE International Conference on Information Technology and e-Services, March 24-26, 2012, Tunisia.

[12] E. Khan, "Information for Everyone using any Phone –…", International Convention on Rehab. Engg. & Assistive Technology in Collaboration with ACM, July 2010, Shanghai, China.

[13] C. N. Hammack et al, "Automated Ontology Learning for a Semantic Web", Dept of CS, Uni. Of Nebraska, Feb 2002.

[14] E. Khan, "Big Data, Natural Language Understanding and Intelligent Agent Based Web", 9th International Conference on Computer Engineering and Applications, Dubai, Feb 22-24, 2015 (Invited Talk & Paper).