Evaluation of Question Answering over the Web

Neil J Wackwitz
Computer Science Department
Saint John's University
Collegeville, USA
njwackwitz@csbsju.edu

Abstract — With millions of internet users turning to search engines for everyday answers to simple questions it's important that one knows whether the information is accurate. This paper will focus and discuss the different methods of natural language processing (NLP) and what website/software is most accurate. To solve this, the paper will explore and analyze some of the most popular websites and software's available to the everyday internet user. The questions we ask will remain the same for each of different software's however; we will also explore the issue of sentence structure within the question. We found that Google was the most accurate among the search engines with Bing close behind but it was also the most work to find an answer. With this new knowledge we concluded that when the question to be answered is a simple fact ask.com was the quickest and most accurate for returning immediate answers. However, Google was the fastest and most powerful when it came to more in-depth questions.

I. History

Natural language processing (NLP) is the computerized attempt to analyze text. There are various definitions for NLP but this definition from the Center of Natural Language Processing (CNLP) provides a good general description for the broad range of NLP. CNLP defines NLP as "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [10].

It is important to note the statement 'range of computational techniques' because currently there multiple methods for the analysis of languages. Also, the statement 'Natural occurring texts' needs a bit more analysis. The only requirement here is the text or oral language must be used by humans. Another important aspect of this statement is that the text being analyzed must be natural occurring, i.e. it was not constructed for analysis purposes. The final statement that must be expanded upon is 'for a range of tasks or applications.' NLP's uses can range from Artificial Intelligence (AI), Information Retrieval (IR), Machine Translation (MT), and Question Answering (QA) [10].

NLP's roots can be traced to the early 1950s where researchers began to treat information and text as information [14]. During the 1950s it was not considered NLP but rather Computational Linguistics (CL). The only difference between the two terms is NLP refers to the actual use in the field while CL refers to the tangible science [10]. The first method, a rather simple method for machine language translation was the dictionary look-up method. This early technology aimed to study the differences between two languages. The computer would parse a sentence word by word and translate each word individually. For example the English sentence "I must go home" would be converted into the German phrase "Ich muss nach Hause gehen." This worked well for small uncomplicated sentences where difference in grammar did not apply. However, for more complex translations a better understanding of the language in question is required. For example, when translating "The spirit is willing but the flesh is weak" from English to Russian and back to English the software would output "The vodka is strong but the meat is rotten" [14]. It was clear for early NLP researchers that simple language translation needed more than just a word for word translation. America was not the only country interested in NLP. China also researched machine translation in the 1950s. The Chinese pursued this technology for about 30 years but like the Americans the research slowed down in the late 1960s. They developed Russian to Chinese and English to Chinese translation programs similar to the German to English program described above [9]. On a different note, almost all of these early dictionary lookup programs were written in LISP which was developed in 1956 by John McCarthy [14].

Machine translation lasted for about ten more years before researchers concluded that it had reached a dead end. From this failure researchers started to develop programs that only dealt with one language. This brought forth translating English into formal statements and in 1968. Brobrow students developed a program that translated algebra word problems into a linear equation. The program would take a problem such as "Billy's age is now 4 times Johns age" and translate it into BA= 2*JA [14]. This development would lead researchers to the more modern natural language data-base.

Two examples of early data-base question answering systems are BASEBALL and LUNAR. The BASEBALL system contained a data-base filled with information from all the games played in the American League for one season. A sample questions would be "Who did the Twins play on July

7?" The LUNAR system used the same data-base system but its information pertained to questions about analysis of rock samples gathered from the lunar moon missions. These two systems proved rather accurate when at the 1971 lunar science convention LUNAR was able to answer 90% of the questions correctly [12]. A third interesting program was developed roughly around 1970. NLPQ introduced a more interactive approach to NLP. The NLPQ program would allow the user to input various declarative statements and by asking the user additional questions it was able to respond to questions about the simulation. For example,

> "User: When a vehicle arrives at a station, it leaves there immediately if the length of the line at the pump in the station is not less than 2.
> User: 75 percent of the vehicles are cars and one fourth are trucks.
> User: There is just one pump.
> User: A simulation time of 8 hours is desired.
> System: How often do the vehicles arrive at the station?
> User: The arrivals of vehicles are normally distributed with a mean of 8 minutes."

The program then continues to ask questions to the user if the information provide was insufficient for completing the desired simulation [14]. All the programs described above lacked one important human element, the element of common sense.

One major issue with the early programs that utilized NLP was that researchers were required to create very efficient algorithms for their programs. Even with the introduction of the world's first vector processing supercomputer, the Cray I, in 1971, programs often took a long period of time to come up with an answer. The Cray I was able to sustain 138 million floating-point operations per second (MFLOPS) and burst upwards of 250 MFLOPS. It also weighed in over 10,000 pounds [13]. This is a far cry from modern day computers. Most modern day computers are capable of doing 3.5 Giga flops per second (GFLOPS) running in sequential and even more than that with the introduction of the dual core, quad core, and most recently 8 core personal computers. Today's super computers such as IBM's RoadRunner is being clocked in at 1.105 PFLOPS that is 10^15 floating point operations per second [1]. With the advancements in computational power and the increasing storage capacity of memory NLP is once again becoming a viable field of study and is allowing sites like google.com to query your search results well under one second and retrieve countless documents relating to the search query.

After entering the early 1970s and into the 1980s there was major research and development in computational linguistics [14]. During this time period researchers expanded on their current systems and started to develop more complex Q & A software. More recently web question answering is becoming more popular. Web pages like trueknowledge.com, askjeeves.com, and wiki.answers.com are being developed every day. Web pages like these use NLP and other forms of question answering techniques to answer the user's questions.

II. Technical Underpinnings

In the resent years the number of online education courses has been steadily rising. An online course taken from Stanford incorporates a videotape of the professor's lecture along with a Power Point [4]. One of the most powerful arguments against online schooling is that it leaves the user with all this information but no interaction. In a classroom setting, students are able to ask questions about something they don't understand, but with online courses this is only possible through email or a chat program. However, in recent years advancements in NLP and artificial intelligence (AI) has made it possible for users to interact with a computer. Many types of webpages and software are capable of answering questions posed by the user. Some people even believe that by the year 2020 computers will reach the capability of the human brain [3]. Furthermore, we will develop various questions pertaining to math, history, science, and current events. These questions will test the current capabilities of question answering webpages. Questions will vary from easy to hard and will also get reworded to see if the sentence structure makes any difference in the answers produced.

Automated question answering (QA) is the process of extracting and expressing information that pertains to the user's initial question. QA software would take a question like "What is Facebook" and the computer would generate a response such as "Facebook is a social network that allows people to connect with friends, share photos, and send messages." We will now examine this further and explore the exact technical underpinnings that allow NLP and other QA methods such as pattern matching, information retrieval, and Google to answer questions.

NLP works by translating a submitted question into a database query where it then extracts the information, retrieves the text that is relevant, and attempts to generate a response, that the user can understand, to the original question. QA software that uses NLP can be broken down into two core areas, the linguistic, i.e., "front end" and DB query/response. The linguistics part of the program takes in the input string and parses it into Natural Language (NL). The parsed words are then sent to the semantic interpreter whose job is to analyze the parsed words based on their suffixes or prefixes. For example, in the question "Who won the 1998 Nobel Peace Prize?", the "who" implies that our answer should be a type "person name" and should be linked with the keywords "won", "1998", "Noble", "peace", and "prize" [1]. After being transformed into a logical query from the semantic interpreter the information is translated into a database query by the query generator. From the query generator the information is passed into the data base management system (DBMS) were it runs the query and stores the results. After the DBMS is finished

with the query the queried information is then sent to a response generator that will output a response to the initial question asked by the user.

Another approach used for QA is pattern matching. Like its name implies this approach looks to match some or the entire question in the target text. For example, consider the question "Who is the CEO of Apple?" The program will search through text to find a similar sentence such as "The CEO of Apple is Steve Jobs." Now for a more in depth look at the algorithm used in pattern matching.

The first step is identifying the type of question. For our example the identifier is "Who is." The sentence can now be rewritten as Q\ = the CEO of Apple. Next, similar to the query generator used in NLP, pattern matching uses existing search engines to mine for answers. After the search the program may have multiple different answers to the same question. It then uses a ranking system to assign a weight to the most relevant answer. In our example there is only one relevant answer so the search might return something like "Steve Jobs is the CEO of Apple, which he co-founded in 1976." However, if the question posed was "What is icing" the search would return with information on both hockey and on baked goods, having to choose which is most relevant. The answer matching software then parses the relevant information "Steve Jobs is" and it is sent to the response generator. This will return the user the final answer "Steve Jobs is the CEO of Apple" [1].

The third version of QA is information retrieval. Information retrieval dose not implicitly answer the question the user poses but basically returns documents that pertain to the users initial question [3]. A prime example of this would be search engines such as Google. Both of the other methods use some form of information retrieval, however they take it a step further and actually extract an answer from the relevant documents.

In order to fully understand information retrieval one must understand the technical underpinning of a search engine such as Google. One might find it incomprehensible on how Google displays webpages relevant to your search results in a split second but it really has three basic steps.

First, Google uses googlebots which are a web crawling robot that finds webpages and passes them to the indexer. These robots find pages in two ways. Through a webpage that allows a user to add his or her webpage into Google's system and by crawling through links found on the web. The first method is quite simple; the page has two input boxes, one for the URL of the site you would like to add, and another to input comments and keywords. The second way the googlebots find a webpage is by crawling from one webpage to another via links. Essentially the robot starts on one page and finds all the other pages it links too, adds them to the crawler queue, and traverses them as well. In addition, Google routinely re-crawls webpages to keep their indexes up to date.

When the googlebot finds a new page it gives the indexer the full text of the page and it is then stored in Google's index database. Each index entry, which is stored alphabetically by search term, stores a list of documents and their location which the term appears. Also words such as the, is, on, or, of, how, or why are not stored in the index as well as single digits and letters. These common words do little to narrow the search.

The final step in the Google search is the query processor. Going beyond just searching and displaying results that have similar keywords Google uses Page-rank to prioritize the results. There are over one hundred factors that go into page rank but a few are the size and position of the search term, the popularity of the page, and the proximity to other search terms.

To analyze these different methods we will ask some of the most popular question answering websites different questions. These questions will pertain to history, science, math, and current events. The four websites we will use for this study are as follows, True Knowledge, Google, and Answers.com.

True Knowledge uses NLP and was able to answer 54% of the questions correctly. However, it is in its beta stage and many of the questions are not loaded into its databases yet. One unique aspect of True Knowledge is that unlike some of the other QA sites it only returned the answer to the initial question. Sites like Google, and Answers.com often returned full pages of text that requires the user to dig through and find the answer.

Google also does not disclose its actual algorithms but after analyzing the test results it mainly appears to use information retrieval since it doesn't return a direct answer but rather a link to a different webpage that contains it. Google successfully provide an answer for each question besides the basic algebra question. It answered 96% correctly. However, it did have conflicting results with two questions. When asked who was the first president of the US the top link was a webpage that talked about how John Hanson was the first and George Washington was the first under the US constitution. Again it had conflicting results when asked the question when Columbus discovered America. Here the top link took us to a site that argued that it was Columbus actually discovered America in 1485 not 1492. However, the source is not scholarly.

Bing is similar to Google since it's a search engine it uses mostly information retrieval. Bing was able to answer all but one question giving it a 96% for correct answers. The only reason Bing scored higher than Google was because Bing was able to solve basic algebra problems while Google was only able to do addition, subtraction, division, and multiplication.

Question Table

Q	True Knowledge	Bing	Google	Answers.com
Who was the first US president?	Yes	Yes	Yes	Yes
When did Columbus discover America?	Yes	Yes	Yes	Yes
Why are there 13 stripes on the American flag?	did not understand question	Yes	Yes	Yes
What is July 4th?	Yes	Yes	Yes	Yes
Who invented the light bulb?	Yes	Yes	Yes	Yes
Where were the first Olympics held?	Yes	Yes	Yes	Yes
	did not understand question	Yes	Yes	Yes
Who was the last president of the Soviet Union?	We don't yet have an answer to that question.	Yes	Yes	Yes
What ethnic group was largely responsible for building most of the early railways in the U.S. West?	Yes	Yes	Yes	Yes
How tall was Abraham Lincoln?	We don't yet have an answer to that question.	Yes	Yes	Yes
When was the great patriotic war?	Yes	Yes	Yes	Yes
Who was the 12th US president?	Yes	Yes	Yes	Yes
Who made the first trans Atlantic flight?	Yes	Yes	Yes	Yes
What is the world's smallest independent country?				
What kind of gun was used to kill Lincoln?	We don't yet have an answer to that question.	Yes	Yes	Yes
What was the last battle of the Napoleonic Wars?	We don't yet have an answer to that question.	Yes	Yes	No answer
When the first Burger King Restaurant opened in 1954, how much did a hamburger cost?	We don't yet have an answer to that question.	Yes	Yes	Yes
How old is the oldest human remains?	Yes	Yes	Yes	Yes
How many people live on earth?	Yes	Yes	Yes	Yes
How many people died in World War 2?	Yes	Yes	Yes	Yes
What is the deepest ocean?	Yes	Yes	Yes	Yes
What temperature does water freeze at?	did no understand	Yes	Yes	Yes
What do we call a scientist who studies fossils from dinosaurs?	We don't yet have an answer to that question.	Yes	Yes	Yes
What does the scientific symbol ag mean?	Sorry, I don't know the answer to that question.	Yes	Yes	no answer
Where was Albert Einstein born?	Yes	Yes	Yes	Yes
What was invented by Samuel F. B. Morse in 1837?	Needed a rewording of the question: electrical telegraph (a telegraph that operates using electrical signals)	Yes	Yes	Yes
What is the speed of light	Yes	Yes	Yes	Yes

Figure 1

Answers.com describes their technical underpinnings as a twofold approach. They don't disclose any actual algorithms but they do say they get their information from "Houghton Mifflin, Columbia University Press, Thomson Gale, Britannica, Barron's, Computer Desktop Encyclopedia, MarketWatch, Investopedia, All Media Guide, Who2, AccuWeather and many more." The second thing they disclose is that Answers.com is connected to WikiAnswers. WikiAnswers is a website where people can ask any questions and other humans respond in a forum type setting. After analyzing the test results it is clear that Answers.com is close in accuracy as a well educated Googler mining through text. It was able to answer 76% of the questions correctly. On another note, Answers.com answers questions with concise answers as well as by displaying full paragraphs of text giving the user a quick answer if it knows and linking a paragraph it believes contains the answer if it doesn't.

The test results are summarized in figure 1.

III. Future Trends

Over the past 60 years Natural Language Processing (NLP) has evolved from a simple translator to complex algorithms used to mine through the web to answer a user's question. Now when we look into the future what does NLP and Question Answering (QA) have in store for us? We can expect to see autonomous robots, responsive online learning environment, and more developed QA software in the next 3 to 5 years. However, the pivotal advancements in this field will occur when a computer will reach the capacity of the human brain. Some people believe that by the year 2020 computers will out compute the human brain [3].

Japan has already developed a robot, Wakamaru that is able to recognize up to 10 faces and can comprehend 10,000 words. The robot weights 66 lbs and is expected to cost $14,000. The Wakamaru will be able to record notes, greet people, and remind owners of his or her appointments [6]. The Wakamaru is a great example of a futuristic technology that NLP will play a significant role. Within 5 years the workplace could possibly have these robotic assistants assisting in daily tasks.

It also claims it is capable of monitoring the condition of a sick person [6]. In hospitals these robots could help the medical staff keep track more patients with less human involvement. For example, they could take blood pressure, monitor temperature, or distribute medication. The current technology would not allow much of the above tasks but, the Wakamaru can communicate simple sentences. So currently these robots could be programmed to give information to the patents and be able to answer simple questions he or she has by combining voice recognition and current QA techniques.

As this technology develops there are many applications that will open up. For example, they could be used in hospitals as doctor assistants, in the army as mechanized solders, and even in the house or office as a helping hand. It's possible that these types of robots could surpass the precision of human doctors and work side by side with humans in a hospital setting. The military could develop an army of robots to perform varying tasks from combat to security detail without risking the lives of humans. The common family could have a robotic maid cooking, cleaning, and performing other household tasks putting less strain on the family.

On a more practical note, futuristic QA software could be used with online learning. Currently one and six students (about 3.2 million) who are enrolled in post high school education took an online course [4]. An online course taken from Stanford University incorporates a videotape of the professor's lecture along with the power point he's lecturing from [4]. Stanford's program came a long way from the traditional way of online courses that basically was a website to take tests and download material. However, Stanford's program does lack one thing, interaction. In a classroom setting, students are able to ask questions about something they don't understand but with online courses this is only possible through email or a chat program. As QA software continues to expand and becomes more accurate, students enrolled in an online course will benefit because any questions can be responded to with immediate feedback.

Currently Google's search engine is only capable of answering questions such as "what is global warming? Or who is Jane Fonda?" However, Google is in its early stages of QA and is only accurately capable of answering questions based on geography, information about famous people, and physical facts, such as the depth of an ocean. We can expect to see Google move in the direction of actually answering questions rather than just returning relevant results [8]. As of right now if you use products such as Google Chrome, Gmail, or AdWords Google is cataloging all your search results. They do this to tailor the results to the specific user based on his or hers interests and past searches [15]. It's likely this technology will also become useful in QA. For example, if I ask the QA software "Where is Saint John's University?" There would be two possible answers, Collegeville Minnesota or Queens New York. If the QA software was using the same technology Google uses to predict what type of page the user wants to see, it would see that I made a recent visit to the Saint John's webpage in Collegeville and know to answer the question with the most relevant information (Collegeville Minnesota) based on my previous search trends.

Another possibility for the future of QA is speech recognition. Currently Google is looking into this technology. Peter Norvig, Google's director of research stated, "We wanted speech technology that could serve as an interface for phones and also index audio text. After looking at the existing technology, we decided to build our own. We thought that, having the data and computational resources that we do, we

could help advance the field.[7]" With speech recognition software implemented a user would be able to ask a question or key terms through their microphone and get instant search results. This piece of software is far more valuable for the mobile phone business since it would allow hands free information lookups when used with a headset.

The future of QA and NLP is not clear to the extent of the technology but, it is clear that these technologies will greatly affect our daily lives. The short term future will result in minor advancements in our current technology but the long term future could produce technology seen from a Sci-Fi movie. With mechanized robots helping with tasks varying from the simple household chores to hands free searching, QA will be around in both the far and near futures. With Google's current stake in the market it's their innovative software/services will likely bring the internet and QA to the next level in the next few years.

REFERENCES

[1] Andrenucci, A., & Sneiders, E. (2005). Automated question answering: Review of the main approaches. Paper presented at the *ICITA '05: Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05) Volume 2,* 514-519. Retrieved from http://dx.doi.org/10.1109/ICITA.2005.78

[2] Answers, "How does Answers.com work?", Answers Corporation 2010, www.answers.com

[3] Cao, J., Robles-Flores, J. A., Roussinov, D., & Nunamaker, J.,Jay F. (2005). Automated question answering from lecture videos: NLP vs. pattern matching. Paper presented at the *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 1,* 43.2. Retrieved fromhttp://dx.doi.org/10.1109/HICSS.2005.113

[4] Gregory W. Hislop, "The Inevitability of Teaching Online," *Computer*, vol. 42, no. 12, pp. 94-96, Dec. 2009, doi:10.1109/MC.2009.411

[5] True Knowledge "The True Knowledge Platform", True Knowledge.(2010)
htpp://www.trueknowledge.com/direct-answer

[6] BBC News. (2005). *Domestic robots to debut in Japan*, Retrieved from http://news.bbc.co.uk/go/pr/fr/-/2/hi/asia-pacific/4196052.stm

[7] Kate Greene, (2007), Future *of Search the head of Google Research talks about his group's projects*, retrieved from http://www.technologyreview.com/Biztech/19050/?a=f

[8] Juan Perez (2005), *Google intro Q&A service*, retrieved from http://www.infoworld.com/t/data-management/google-intros-qa-service-511

[9] Zong, C., & Gao, Q. (2008). Chinese R\&D in natural language technology. *IEEE Intelligent Systems, 23*(6), 42-48. Retrieved from http://dx.doi.org/10.1109/MIS.2008.100

[10] The Computer Language Company Inc. (2010). *Definition of: Computational linguistics.*http://www.pcmag.com/encyclopedia_term/0,2542,t=computational+linguistics&i=40134,00.asp

[11] Narayan, S. (2009). Supercomputers: Past, present and the future. *Crossroads, 15*(4), 7-10. Retrieved from http://doi.acm.org.ezproxy.csbsju.edu/10.1145/1558897.1558900

[12] Moll'a Diego, & Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Comput.Linguist., 33*(1), 41-61. Retrieved from http://dx.doi.org/10.1162/coli.2007.33.1.41

[13] Petersen, W. P. (1983). Vector fortran for numerical problems on CRAY-1. *Commun.ACM, 26*(11), 1008-1021. Retrieved from http://doi.acm.org.ezproxy.csbsju.edu/10.1145/182.358469

[14] Lehnert, W. G., & Ringle, M. H. (1982). Strategies for natural language processing

[15] Buresh S. (2007), *Current and Future Search Trends: What the Top Interent Search engines Are Doing*, Retrieved from http://www.searchengineguide.com/scott-buresh/current-and-future-search-trends-what-th.php