INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & STUDIES

IJIRS

Survey of Most Powerful Language Software's

ISSN 2319-9725

Piyush Pratap Singh

Asst. Professor Information technology

Mahatma Gandhi International Hindi University

Abstract: *Natural language processing NLP, a branch of Artificial Intelligence AI is developing very rapidly in all R&D organization, commercial organization and software companies. It is known that NLP is a vast field of research which has many sub fields as natural language understanding, parsing, machine translation (MT), speech recognition, grammatical analysis and many more. Different organization having different priorities are developing their respective techniques and software's if we try to count the working software and techniques around the world on NLP the tally goes in thousands. This paper compares the technology of three most powerful NLP software's and makes review of them.*

Keywords*: AI, NLP, MT, MS*

1. Introduction:

NLP is the fastest developing field of AI because the power of intelligence hidden behind the system, in the age of information technology where all types of communication is being done by machines, it is essential for the machine to efficiently represent the language so that information can be retrieved from the system, globalization further push the system to translate and understand the meaning of different languages. The most recent and advance technology is speech recognition and interpretation if achieved will change the entire technology because it will recognize speech, change speech to text and interpret the text for desired task example: translate to destination language.

NLP software's, researches and techniques to implement them are designed every day in tons but most of them not even appear to public as some are inefficient, some crash and others do not sustain in the software market as they are outdated even before their launch. The reason behind this is the software tycoons having at least 500 highly qualified engineers with a thousand member team having team leaders, system analyst, risk managers, error managers, testing team, management team etc. working on a single software and so they out rank the others as they use the most advance technology with no fund limitations, resources or man power crisis in an excellent infrastructure.

The most powerful NLP software working in public domain in recent are Microsoft Word grammar checker by Microsoft (MS) technologies, Google translate by Google and most recent and advance is Apple iOS 6 Siri by Apple all the above belongs to NLP, but the technology they are using is different. There are many more strong language software made by NASA, Russian Defense, ISRO and nearly all the countries but we do not have their access, and the rest available software are far behind the above due to the discussed reasons.

Microsoft, Google and Apple all are profit making commercial companies so we do not have the access to the design and algorithms still we try to understand the working and limitations .

2. Microsoft Word Grammar Checker:

NLP has been developed in Microsoft before Google and Apple, Microsoft NLP system is called "NLPWin" [1] that has been developed at Microsoft Research facilities over the last 15 years. NLPWin consists of the following major components

Lexical Processing includes tokenization, word segmentation, morphological analysis of words, the identification of multi-word entries and dictionary lookup. Syntactic Processing refers to the parsing of sentences to produce syntactic descriptions of the various segments therein. Logical Form (LF) specifies semantic relationships among the various segments of a sentence. Generation produces a sentence string from an LF. Alignment is a part example-based, part statistical process for extracting corresponding LF segments of two languages based on bilingual corpus data; it is used for their Machine Translation (MT) system [1].

Their LF represents the semantic relationships of the arguments within a sentence. For instance, Figure 1 illustrates the LF of the sentence, "At school, John eats rice every day."

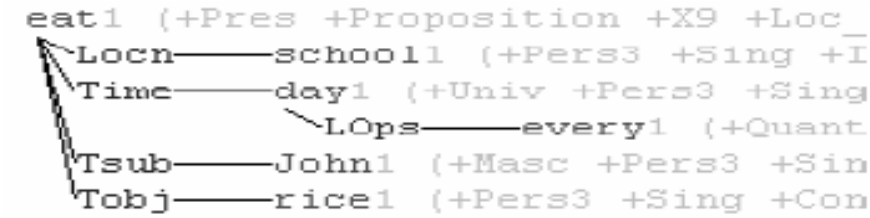

Figure 1: An LF produced by NLPWin

From LF, Generation can automatically produce wh-questions (i.e., "who/what/when/where/why/how" questions) for a given text with increasing levels of complexity. From the above input, for instance, they can generate the following questions-answer pairs:[1]

```
==========================================
<Q>At school, what does John eat every day?</Q>
<A>Rice</A>
==========================================
<Q>At school, who eats rice every day?</Q>
<A>John</A>
==========================================
<Q>Where does John eat rice every day?</Q>
<A>At school</A>
```

Figure 2: QAs generated from the LF in Figure 1

NLPWin contains rich lexical information. Their generation component can utilize lexical information to create grammar-based exercises automatically. For instance, from the simple sentence, "The man eats rice." they can generate numerous variations of that sentence, including those below. [1]

Input: The man eats rice.
Present Passive => Rice is eaten by the man.
Present Perfective=> The man has eaten rice.
Present Progressive=> The man is eating rice.
Past => The man ate rice.
etc.

Figure 3: Verb Inflection Exercise

Input: The man eats rice.
Plural Subject => The men eat rice.
Plural Object => The man eats rice.

Figure 4: Noun Inflection Exercise

Word:	**Part-of-speech**
rice =>	NOUN
eat =>	VERB
man =>	NOUN
the =>	ADJ

Figure 5: Part-of-speech Exercise

The Grammar checker software was first found in UNIX operating system way back in 1970, was basically a programs that checked for punctuation and style inconsistencies. In 90's Microsoft word 97 was launched with a grammar checker with a controversy that weather it will help or hinder students, word 97's Grammar Checker catches some errors reliably (e.g. subject verb agreement errors) and others at least occasionally (e.g. comma and capitalization errors). However, it does not catch most pronoun or modifier errors [2].

2.1. Technical Issue:

The process of MS is based on simple pattern matching. The heart of the program is a list of many hundreds or thousands of phrases that are considered poor writing by many experts. The list of suspect phrases included alternative wording for each phrase. The checking program would simply break text into sentences, check for any matches in the phrase dictionary, and flag suspect phrases and show an alternative. These programs could also perform some mechanical checks. For example, they would typically flag doubled words, doubled punctuation, some capitalization errors and other simple mechanical mistakes.

Patricia J. McAlexander in his thorough study "Checking the grammar checker: Integrating grammar instruction with writing" [3] of nearly all possible phrases scores the grammar checker as

THE GRAMMAR CHECKER'S SCORES [3]

ACAE Project

FRAGMENTS	71%
COMMA SPLICES	75%
FUSED SENTENCES	0%
COMMAS IN LISTS	67%
COMMAS IN COMPOUND SENTENCES	0%
COMMAS WITH INTERRUPTERS	40% (can do which and that clause)
COMMAS WITH INTRODUCTORY ELEMENTS	50%
QUOTATIONS (can do periods and commas inside quotation marks if set to do so)	
COLON	50%
DASH	0%
SUBJECT-VERB AGREEMENT	83%
PRONOUN AGREEMENT	50%
ACTIVE-PASSIVE VOICE will mark passive voices every time, but you may want to use the passive!	
APOSTROPHES	60%
PARALLELISM	25%
MODIFIERS	0%
PRONOUN REFERENCE	0%

Table 1: The Grammar Checker's Scores

Since the study Word has made some more improvements points

3. Google Translate:

The R&D at Google is developing at very rapid rate because users of Google are multiplying in quick time. Google search is the mostly used application of Google.

Google's focus on innovation, its services model, its large user community, its talented team, and the evolutionary nature of Computer Science research has led Google to a "Hybrid Research Model."[4] They have started multi-year, large systems efforts (e.g., Google Translate, Chrome, and Google Health) that have important research components.

Google Translate is a free translation service that provides instant translations between 57 different languages. [5]

Google Translate generates a translation by looking for patterns in hundreds of millions of documents to help decide on the best translation. By detecting patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "statistical machine translation". It can translate text, documents, web pages etc. English to Hindi Machine Translation system (http://translate.google.com/), in 2007, Franz- Josef Och applied the statistical machine translation approach for Google Translate from English to other Languages and vice versa, thus statistical machine translation approach for identification of idioms is proposed. Many online machine translation systems are available on internet as no single translation engine will be consistently most effective for all pairs of languages and text conditions. As further we use Google Translate system for translating English Text to Hindi. The accuracy is good enough to word understand the translated text, but not perfect [6] for example English: It's raining like cats and dogs, Hindi translation By Google: अपनी बिल्लियों और कुत्तों की तरह बारिश हो रही and how to translate the following आग में घी का काम करना which means "Add fuel to fire"

In a Methods Research Report "Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages" Prepared for: Agency for Healthcare Research and Quality U.S. Department of Health and Human Services Prepared by Tufts Evidence-based Practice Center, Tufts Medical Center: Boston, MA assess the accuracy of Google translate in different languages and concluded the following result.[7]

Confidence	Chinese, Percent Accurate	French,* Percent Accurate	German, Percent Accurate	Japanese, Percent Accurate	Spanish, Percent Accurate	Overall,* Percent Accurate
Strong	73%	94%	78%	66%	80%	79%
Moderate	78%	76%	77%	69%	83%	76%
Little	88%	67%	74%	75%	74%	76%

* 1 extractor did not rate confidence level for 1 article.

Table 2: Association between extractors' confidence in accuracy of translation and their extraction accuracy

Bhojraj Singh Dhakar[8], also in their paper compared Google and Bing translators on four inputs News Paragraph, Technology Paragraph, Medical Paragraph, Official Documents on following categories Missing Words, Word Order, Incorrect Words, Unknown Words, Punctuation errors. Google was found better in all the four inputs.

4. Apple Siri:

A number of tech companies - including Google, Microsoft and IBM - are working on voice technology. But Apple's introduction of Siri, combined with their marketing, has really had an impact in creating consumer expectations.

Voice technology is one of most advanced fields of NLP and in recent years, Google has introduced voice-activated navigation, search and translation apps based on technology that it developed internally and through acquisitions. But with the debut of Apple's Siri last year, the tech rivals are now engaged in a high-profile voice-technology arms race. Microsoft has also added voice capabilities to its smart phone software as it vies for a share of the mobile market now dominated by Apple and Android. In addition, Microsoft has introduced voice commands for its Bing search engine and Xbox entertainment console

Meanwhile, Ford and other automakers are incorporating voice technology into navigation systems, climate systems and music players.

Siri first listen your voice that is speech recognition than convert it to text process command or instruction to do what is desired by user technically we can say that a guided dialog to domain and task model to the web services and API's .

The current version of Siri is not as efficient as the company projected it you will find it on using as the first step voice recognition is 30% efficient, many phrases that i had used from one month like a simple one call piyush it will call 3 times if I will say it 10 times, and any phrase which is more than 5 sec is more hard for the software to interpret. Voice recognition is what Siri does, but those words alone don't reveal how the system actually gets your words right when you say.

4.1. Technique Of Siri:

The sounds of your speech is immediately encoded into a compact digital form that preserves its information.The signal from your connected phone is relayed wirelessly through a nearby cell tower and through a series of land lines back to your Internet Service Provider where it then communicated with a server in the cloud, loaded with a series of models honed to comprehend language. Simultaneously, your speech is evaluated locally, on your device. A recognizer installed on your phone communicates with that server in the cloud to test whether the command can be best handled locally -- such as if you had asked it to play a song on your phone -- or if it must connect to the network for further assistance. (If the local recognizer finds its model sufficient to process your speech, it tells the server in the cloud that it is no longer needed. The server compares your speech against a statistical model to estimate, based on the sounds you spoke and the order in which you spoke them, what letters might constitute it. (At the same time, the local recognizer compares your speech to a reduced version of that statistical model.) For both, the highest-probability estimates get the go-ahead.[10]

Based on these opinions, your speech -- now understood as a series of vowels and consonants -- is then run through a language model, which estimates the words that your speech is comprised of. Given a sufficient level of confidence, the computer then creates a candidate list of interpretations for what the sequence of words in your speech might mean. If there is enough confidence in this result, -- the computer determines that your intent is to call, piyush is your addressee (and therefore his contact information should be pulled from your phone's contact list) -- your dialing box magically appears on screen, no hands necessary. If your speech is too ambiguous at any point during the process, the computers will defer to you, the user: did you mean payush, or pratush? Or many times will say that: I don’t understand you, etc.

Siri can be summarized as

4.1.1. Doing For Us (User):

Service API's, Data Feeds, Faceted Search Domain

In future: Auth Standards, SW and Data commons, Recommendation services

4.1.2. Getting What We Say:

Location, Speech, and Semantic NLP, Date/Time, Conversational UI

In future: Linguistic NLP, Social contexts. [11]

5. Result:

From the above discussion it is clear that due to large infrastructure Google, Microsoft, Apple, IBM will rule the NLP sector for times to come. Siri, Google translate and MS Grammar checker are the software of future.

6. Conclusion:

Development in NLP will be on the highest priority as the tech rivals are now engaged in a high-profile NLP arms race. The user enjoys the optimum result as ultimately he will have a cost effective technology.

References:

1. Schwartz, Lee, Aikawa Takako, & Pahud Michel "Dynamic Language Learning Tools." Natural Language Processing & Learning Sciences and Technology, Microsoft Research, Redmond, WA, USA InSTIL/ICALL2004 Symposium on Computer-Assisted Language Learning, Venice, Italy, June 17-19 , 2004.
2. Caroline Haist, Canadore College, North Bay, Ontario "An Evaluation of Microsoft Word 97's Grammar Checker" Book chapter January 2000.
3. Patricia J. McAlexander "Checking the grammar checker: Integrating grammar instruction with writing " journal of Basic Writing, Vol. 19, No. 2, 2000
4. Alfred Spector, Peter Norvig, Slav Petrov "Google's Hybrid Approach to Research" Google Inc. Communications of the ACM Volume 55 Issue 7, July 2012Pages 34-37 Doi :10.1145/2209249.2209262
5. http://translate.google.com.
6. Monika Gaule , Gurpreet Singh Josan "Machine Translation of Idioms from English to Hindi" (ijceronline.com) Vol. 2 Issue. 6 Issn 2250-3005(online) October| 2012
7. Ethan M. Balk, Mei Chung, Minghua L. Chen, Thomas A. Trikalinos, Lina Kong Win Chang "Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages" AHRQ Publication No. 12(13)-EHC145-EF January 2013
8. Bhojraj Singh Dhakar, Sitesh Kumar Sinha, Krishna Kumar Pandey "A survey otf translation quality of English to Hindi Online translation systems (Google and Bing) "International Journal of Scientific and Research Publications, Volume 3, Issue 1, January 2013 ISSN 2250-3153
9. http://office.microsoft.com/en-in/word-help/check-spelling-and-grammar-HP010117963.aspx
10. http://www.zdnet.com/ applesiri
11. www.siri.com/presentation
12. Sitender, Seema Bawa "Survey of Indian Machine Translation Systems" IJCST Vol. 3, Issssue 1, Jan. - March 2012 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
13. ALEX VERNON "Computerized Grammar Checkers 2000:Capabilities, Limitations, and Pedagogical Possibilities" Computers and Composition 17, 329–349 (2000) ISSN 8755-4615
14. http://seattlepi.nwsource.com/business/217802_grammar28.asp