Intelligent News Reader

KLTN Perera #, PPNV Kumara [2], and B Hettige [3]

[1,2]Department of Information Technology, Faculty of Computing
General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka.
[3]Department of Computer Science, Faculty of Computing
General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka.

KLTN Perera; <starnp8@gmail.com>
[2] N Pathirage; <nandana_pat@yahoo.com>
[3] B Hettige; <Buddhitha@yahoo.com>

Abstract*— News is one of the most essential requirements from the beginning of the history, it is anything that is exciting, that narrates to what's happening in the world, what's happening in areas of the nation that would be of interest to the particular audience. With the invention of media people get to know about the current situation of the country as well as the whole world through it. But because of the complex routines of life they are not able to allocate much time to listen news. Internet has transformed the boundaries of news with fast and easy access. But still people bothered about the validity of the news they heard. To become a good news it should be clear, precise and brief.*

This paper reports the design and implementation of the "Intelligent News Reader" application which can be known as a great solution to the busy scheduled users concerning their thirst of news. Basically this product is based on the natural language processing, text to speech conversion, which converts the characters in words into audio format that sounds just as the words are spoken. News which is needed for the software is taken from the prominent news websites with the use of web parsers. The news taken from the websites are finally converted into a meaningful format, which can be easily captured by the users.

Keywords*—* **Web Scrapping, HTML parsing, Text to Speech Conversion, Text comparison algorithm**

I. INTRODUCTION

News has become the most important source to get a knowledge regarding the day to day events happening around the globe. News travels through many different media based on printing, broadcasting, postal systems and communication. Since majority of the society are really busy with their routines activities, they cannot allocate much time for listen or observe news broadcast on television, Radio Channels or News Papers. Consequently in the present society most of the people are not aware about the current situation in the world. Because of that social media transformed the way consumers receive and participate in news. Gradually they lured people towards social media sites, gossip sites, online reading sites which people consider as unreliable sources. (Alejandro, 2010)Since most of the people deal with their laptops and handheld device all the time, and if they are able to listen for those news while engage in their day to day works it will be a wonderful experience for them. Intelligent News Reader is an automated solution for the busy scheduled users regarding the day to day events around the globe. It affords news for its users from reliable sources (like prominent news websites) by extracting news in an accurate format and speak them out to its users in a clear and precise manner. A user can separately listen to the headlines and news in detail by using this application and he or she can activate this program in any time period of day and listen to the news while doing his or her routine activities.

There are no similar applications to this solution. Currently there are text readers built from several languages. They are capable of reading a given text to the listener. For that speech synthesizers are used. Several text readers are produced from java, python and C#. Html parsers are needed for the proposed application to grab the news from the prominent websites. DOM-based Content Extraction of HTML Documents is an early research done related to the topic. Abstraction method of text summarization, which is based on natural language generation is used to develop this application.(Michele Banko, n.d.) The Columbia Newsblaster1 system has been providing summaries of topically clustered news daily and it aid daily news browsing by providing an automatic, user-friendly admittance to important news topics, along with summaries and links to the original articles for further information.

This paper reports the design and implementation of "Intelligent News Reader which is designed to the busy scheduled users. The design architecture and technology methods are also given in the paper.

The rest of the paper is organized as follows. Section 2 reports brief summary of the existing frameworks and systems. Section 3 gives on the technology behind this application. Then section 4 gives Design and implementation of the Intelligent News Reader with a brief description of each module and section 5 depicts how application works as defined. And also section 6 presents about the results of evaluation. Finally section 7 gives conclusion and further works of the project.

II. CURRENT PRACTICES

Currently there are no similar applications to this product, but there are applications created by several researchers by using the key technologies used to develop this application. Here presents about the key findings which are already available in the field of technology and technologies which inspired to carry out this application.

A. Text to Speech Conversion

Text to speech conversion is used to create models of the human vocal tract to generate a synthetic human voice output conforming to input text. Text-To-Speech system processes are usually different than live human speech invention. Text-To-Speech conversion contains of two parts; namely the Natural Language Processing (NLP) unit and the Digital Signal Processing (DSP) unit. Natural Language Processing unit consider phonetization and pitch along with rhythm and it outputs a phonetic transcript of the input text. The Digital Signal Processing unit converts the phonetic transcript it obtains into machine speech. (World Congress on Engineering et al., 2014)

It should contain text analyzer which contains pre-processing block which converts numbers, abbreviations and acronyms into full text when needed. The morphological analysis block which classifies each term in the sentence presence analyzed into thinkable parts of speech, like word's spelling. The contextual analysis module which modernizes the list of conceivable parts of speech of words in sentences and finally the syntactic-prosodic parser track down the text structure. And Phonetization model which helps to phonetic transcription of incoming text and Prosody generation concentration on specific parts of a sentence, such as emphasis laid on an exact syllable, thus attributing special position or difference to that part of the sentence.

Rhythm is a significant factor that marks the synthesized speech of a TTS system more natural and comprehensible, the prosodic configuration offers essential information for the prosody generation model to generate effects in synthesized speech. Numerous TTS systems are established based on the principle, corpus-based speech generation. It is very common for its high feature and natural speech output.

When relating to text readers, there is a German text-to-speech synthesis system Mary, which was established as a research and development tool. It contains with modular design and XML-based system-internal data illustration. However, as the MARY system uses an XML based data representation. (Schröder and Trouvain, 2003) System receives both plain text input and input marked up for speech synthesis. The input markup language, currently SABLE and the W3C draft version of SSML, is interpreted by this module into the system-internal Mary XML format, upon which successive modules will operate. The understanding indications uttered in the input markup are reflected as supplements to the modules' text-to-speech analysis of the input. Each module enhances new or more comprehensive information. The tokeniser cuts the text into tokens, i.e. numbers, words, special characters and punctuation marks. It uses a set of rules firm through corpus analysis to label the meaning of dots based on the nearby context.

B. html parsers and Web Scrapping

html parsers are needed for the proposed application to grab the news from the prominent websites. DOM-based Content Extraction of HTML Documents is an early research done related to the topic. In directive to examine a web page for content extraction, the page is first passed through an HTML parser that modifies the HTML and creates a Document Object Model tree illustration of the web page.

Scrappy is intended to scrape web content from sites that are collected of many pages of comparable semantic arrangement.(Amir Ghazvinian1, n.d.) The system is employed as a Firefox browser extension, and works in three main steps to scrape web data. First, a user steers to a page that he would like to scrape and makes a model for the content that he would like from that page. Then, the user chooses a set of links that point to pages matching the content template definite by the user. First phase in obtaining data from the web by Scrappy is to make a template that can be functional to scrap data from multiple pages of comparable structure. The user steers to a page that comprises links to a set of pages corresponding the template quantified in the previous step. Here, the user can choice the links that he needs to scrape simply by click on them. When the user hovers over a link in link assortment mode, Scrappy highpoints a set of similar links. Clicking on any of these links will choice and highlight all of them for scraping, which saves the user time in agreeing links for scraping. Once the user has quantified a content template and has designated a set of links to crawl, he simply selects an output. That is the theory behind the Scrappy which is developed by the Stanford University.

C. Text Comparison and Processing Method

Since there are same news from different news websites, it is essential to remove the duplicates. For that text comparison algorithms have been used. Those algorithms are called "String metric algorithms". Such algorithms are used to find the word patterns with the variances like insertions, deletions or replacements.(Jokinen et al., 1988)For the purpose of learning the differences between words there are several algorithms used in the computer science field. These algorithms are used in dynamic programming for the purpose of having an optimized result. (Pandiselvam.P and Marimuthu.T, n.d.)(Cohen et al., 2003)The algorithms can be known as Hamming distance algorithm, Levenshtein algorithm, Smith Waterman algorithm, Boyer –moore algorithm and brute force algorithm.

Concerning the advantages and disadvantages of several algorithms "Levenshtien algorithm "has been used to build this application. It calculates minimum number of character edits required to edit two words. (Cheapest way to transform one string to another.) This transformation can be an insertion, deletion or replacement of a character. This algorithm also can be known as "Edit distance algorithm".(Van der Loo, 2014)

When concerning about the usage of this algorithm, spell checking applications, correction structures for optical character recognition, and software to support natural language translation based on translation memory, DNA analysis, and plagiarism detection can be identified. The Levenshtein distance can also be calculated between two lengthier strings, but the cost to calculate it, which is approximately proportional to the product of the two string sizes, makes this unrealistic. Therefore, when used to aid in fuzzy string searching in solicitations such as record linkage, the matched strings are commonly short to help increase speed of evaluations.

To find the similarity between the tokens "Bipartite graph" was used. For that "Assignment problem" is used.(Kuhn, 2010) Bipartite graph is a graph whose vertices can be separated into two disjoint sets and every edge links a vertex in to one in and are generally called the parts of the graph. The bipartite graph configuration can be used to seizure a relationship between two types of objects where the dissimilarity between the types of objects is significant. In order to solve the assignment problem using bipartite graph, Hungarian method is used. From Hungarian method, a user can calculate the maximum weight of bipartite match.

As mentioned earlier, there are several models developed by several researchers using the technologies which used in this application. But there is no similar application like this "Intelligent News Reader". By evaluating and considering the limitations and the advantages of those implementations "Intelligent News Reader" is developed. Below section put forward about the technologies and theoretical advancements which have been used to develop this product.

III. TECHNOLOGY BEHIND THE APPLICATION

This section depicts about the theory and technologies lies behind this application.

A.Text to Speech Conversion

Text to Speech technology is the method that used to convert the normal text format into an audio format. Here the sentence which needs to be read should be properly analyzed. This analyze should cover whether there are nouns, phrases in the sentence, tense of the sentence and also the beginning and the end of the sentence.

When concerning about this "Speech Synthesis" process, there are three major steps related to it. First thing is "Text Normalization". Here in text normalization it identifies the phrases and the parts in the sentence. The second process is "Linguistic Analysis.(Swetha and Anuradha, 2013) Here it recognizes the process or the stream of the sentence. Normally there are alterations between written and spoken forms of a language, and these variances can lead to indeterminacy or indistinctness in the pronunciation of written words. Therefore the step should be properly covered. The third step is the "Prosody Generation". Prosody is the set of structures of speech output that comprises the pitch (melody or intonation), the rhythm, the pausing, the speaking rate, the weight on words and many other features. Producing human-like prosody is significant for creating the sound of speech more natural and for properly conveying the meaning of spoken linguistic.(Dennis H. Klatt, 1987)

Text to speech applications always read sentences as single, context-independent substances therefore it is needed to thoroughly consider about the spellings and the text formatting in the sentence. Commas are used to consider the pauses in the text and it will create a significant impact to the synthesis speech. Without commas, Text to speech sounds too fast and abnormal. There should be intervals between the each and every sentences to become more natural. Therefore it is really good to consider about the correct punctuations in the application.

B. Use of Web Scrapping and xml Document

Most of the Webpages are designed using text based mark-up languages like HTML and XHTML. To extract the news from the websites, reaching to its code level is essential. According to the requirement, grabbing the headlines and the detailed news is the major functionality behind the application. Therefore separate XPath queries have been written to identify the paths to the relevant news websites.

As per the requirement the application should be able to format the news into a meaningful manner and the repeating groups and unnesseray points should be removed, for this purpose the application loads the news into a separate XML document, in order to manipulate the proper outcome from the news. This xml document will be helpful to analyze the news into a proper format and gives a meaningful output to the users.

C. Text Comparison and Processing

The news required for this application are taken from the prominent news websites. Therefore the contents of those websites can become similar sometimes. If that kind of situation occur in this application, it will become a nuisance for the users and it will reduce the quality of the application. Hence it is essential to remove the duplicative content from the application. To remove these duplicate content several string metric algorithms have been used. (Gao et al., 2010)

As the first method "Levenshtien Algorithm" is used. Here it consider about the minimum number of character edits required to transform a one word to another.(Eric Sven Ristad, n.d.) Here using this algorithm application calculates this edit distance and greater the levenshtien distance means strings are different. Since it is essential to find the stings with smaller differences, this is the most suitable algorithm. Since the news are created using sentences, it is needed to divide the each sentence to peace of information which is technically considered as "tokens". The characters which separates the words are recognized as delimiters (space, commas etc.) and to identify the delimiters "Regular Expressions are used.

Normally sentences can be considered as a list of tokens. Therefore to identify the similarities between those sentences "Hungarian method" was used. This method was used as reducing the assignment problem from a bipartite graph. The usage of Hungarian method is to identify the total weight of this bipartite match. From this the needed optimized output can be achieved.

Below section depicts about the design and implementation details of the "Intelligent News Reader" application.

IV. DESIGN AND IMPLEMENTATION OF INTELLIGENT NEWS READER

This application is developed as a standalone application, and application is developed using asp.net with the C#. The application will be able to browse the web when "Go" button is pressed by the user. The application will automatically visit the prominent news websites which are pre-defined by the user. At that time the application will extracts news from the visited websites and categorized them into Headlines and News-in-details into an xml document. After that the extracted news will be processed according to a proper format. Finally the application will speak-out those news to the user in clear and precise manner. The user of the application will be able to change the volume and frequency of the application according to the preference. The application will consist with user-friendly interfaces which can be easily controlled by the user.

When concerning about the process of the application, the application can be segregated into three basic functions. They are Select the appropriate news website, Extract the news from the website, Order the news in an appropriate manner, process the news and finally Read out the news to the user.

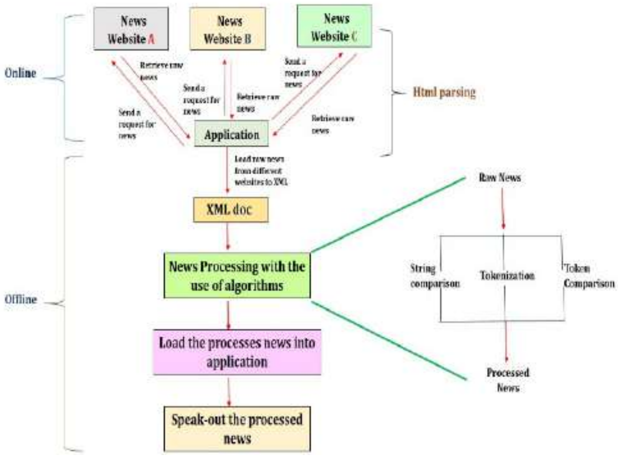

Figure 1: Process Diagram of the Application

A. *Extracting News*

This is the first module of this application. Only for this module a device with internet access is required. When user starts the application and press "Go" button, application will go to the news websites. Here the application will visit to the default news websites. In addition to that user can customize the news websites as his or her preference.

B. Text Processing and Comparison

Text analysis and summarization is the second major module in the application. This process is happening as an internal module, when user press "Go" button. As it is mentioned in the previous section, when user press "Go" button application will automatically routes to the news websites and retrieve news from it. Those news are loaded into an xml document. That xml document is used to do the text analysis process.

Since application took news from several news websites, there might be duplicative news in it, if they are available in the application, it will reduce the productivity of the application. Therefore it is essential to remove those content from the application. Therefore when the match score value is become high, it means the sentences are similar enough, it means they are duplicative. Because of that it is essential to remove a one news from it. Here the application will remove a one news content when there are same news in different news website. When the words matching count is become 50% or more than 50% this functionality works that is the process where it users the text processing part.

Using the content generated in the xml document, it analyses the similar content and remove them. After that it keeps one news content, which is more descriptive and can be categorized as the best one. The news will automatically order according to the sequence of headline and news in details.

C. *Presenting News*

This is the final output of the application and this is the output which can be seen and listen by the user. When user selects the "Play" button. Application will start speaking

V. HOW APPLICATION WORKS

This application is developed using visual C# language as a windows form application. The only difference lies here is, it can access internet when necessary and grab the news. Since this is an application to use in a daily basis, it should have attractive graphical user interfaces. Therefore application developed with several animations and eye catching theme. It consists with three major modules and finally output is displayed as a single application.

The interfaces are developed to have a high user experience and this is really a user friendly application. When user press "Go" button, application is going to specified news websites and retrieve the news from those news websites and load them into an xml document. The news processing part also happening in here. If user want to add more sites to the application, user can select the settings button. Using the settings button user can add or remove news websites according to the preference. And also can select whether the application needs to read the news as headlines or detailed news or both.

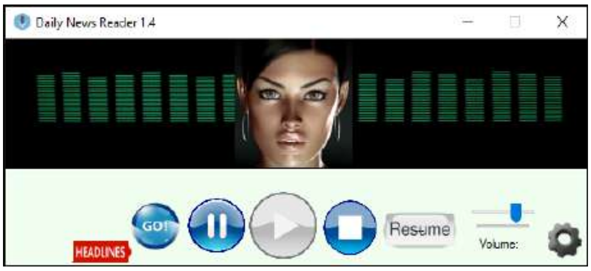

Figure 2: Interface of application when user selects play button

Button integration of this application also need to concern, all the related buttons are activated when they are necessary to activate. Other time the buttons remains disabled.

Using the button "Headlines" user can listen to the headlines only. And also using the volume controller user can adjust the volume according to the preference.
Another important thing in this application is the presenting style. It presents news in a very good pronunciation method and it is presenting style is very similar to television news presenter .Therefore the application will become a realistic experience to its user, though this is an artificially generated application.

VI. RESULTS

The accuracy evaluation was conducted in order to find whether the match score of "Intelligent News Reader " application is similar to the match score given by the human evaluator. This criteria is really essential to remove the duplicate news coming from several news websites.

Accuracy testing was conducted by the use of hundred pairs of sentences. Then the each pair of sentences were dropped to the demo application which was created for evaluation purpose and compare the similarity by giving a match score value. After that each pair of sentences were given to a human being to consider the similarity of those sentences. The results match scores of the application and human evaluator were compared and analyzed properly. The results of both evaluations are as follows.

Table 1: Results Coming from Human Evaluator and from the Application

Couple of sentences	Match score by Human	Match Score by application
Couple 1	71-80	69.23%
Couple 2	11-20	35.29%
Couple 3	71-80	71.6%
Couple 4	91-100	94.8%
Couple 5	61-70	56.09%

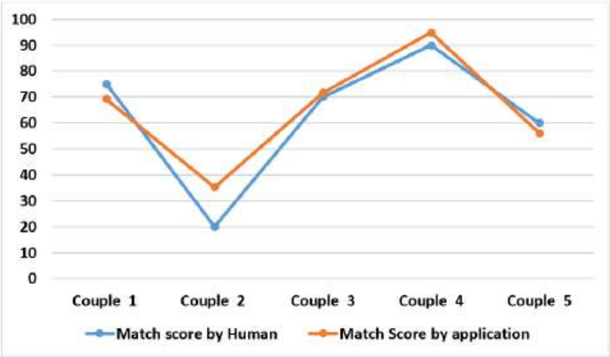

Figure 3: Difference between Results Coming from Human Evaluator and from the Application

Here the test results of human evaluator were appeared similarly as given by the application. Therefore the selected criteria for removing the duplicates (lesser than fifty) was recognized as a correct value.

VII. CONCLUSION AND FURTHER WORKS

This paper has reported the design and implementation of the “Intelligent News Reader”. Further it has provided the evaluation results of the application. By analyzing the results, it is clear that Intelligent News Reader application completely meets its expected goals. As the future enhancement this application will be developed for Sinhala language.

ACKNOWLEDGEMENT

I acknowledge Mr. Nandana Pathirage and Mr. Buddhitha Hettige for encouraging me to write this paper, and also the lecturers who inspire me to write this. The author bears all the responsibility for any misunderstandings and/or errors therein.

References

Alejandro, J., 2010. Journalism in the age of social media. Reuters Inst. Fellowsh. Pap. Univ. Oxf. 2009–2010.

Amir Ghazvinian1, n.d. Scrappy: Simple Web Scraping.

Cohen, W., Ravikumar, P., Fienberg, S., 2003. A comparison of string metrics for matching names and records, in: Kdd Workshop on Data Cleaning and Object Consolidation. pp. 73–78.

Dennis H. Klatt, 1987. Review of text-to-speech conversion for English.

Eric Sven Ristad, n.d. Learning String Edit Distance1.

Gao, X., Xiao, B., Tao, D., Li, X., 2010. A survey of graph edit distance. Pattern Anal. Appl. 13, 113–129. doi:10.1007/s10044-008-0141-y

Jokinen, P., Tarhio, J., Ukkonen, E., 1988. A Comparison of Approximate String Matching Algorithms. SOFTWARE—PRACTICE Exp. 1, 1–4.

Kuhn, H.W., 2010. The Hungarian Method for the Assignment Problem, in: Jünger, M., Liebling, T.M., Naddef, D., Nemhauser, G.L., Pulleyblank, W.R., Reinelt, G., Rinaldi, G., Wolsey, L.A. (Eds.), 50 Years of Integer Programming 1958-2008. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 29–47.

Michele Banko, n.d. Open Information Extraction from theWeb.

Pandiselvam.P, Marimuthu.T, n.d. A COMPARATIVE STUDY ON STRING MATCHING ALGORITHMS OF BIOLOGICAL SEQUENCES.

Schröder, M., Trouvain, J., 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. Int. J. Speech Technol. 6, 365–377.

Swetha, N., Anuradha, K., 2013. Text to speech conversion. Int J Adv Trends Comput Sci Eng 2, 269–278.

Van der Loo, M.P., 2014. The stringdist package for approximate string matching. The R.

World Congress on Engineering, Ao, S.I., Gelman, L., Hukins, D.W.L., Hunter, A., Korsunsky, A., International Association of Engineers (Eds.), 2014. World Congress on Engineering: WCE 2014 : 2-4 July, 2014, Imperial College London, London, U.K.