American Journal of Information Science and Computer Engineering
Vol. 7, No. 3, 2021, pp. 36-51
http://www.aiscience.org/journal/ajisce
ISSN: 2381-7488 (Print); ISSN: 2381-7496 (Online)

On the Development of Machine Learning Algorithms for Information Extraction of Structured Academic Data from Unstructured Web Documents

Joshua Babatunde Agbogun[1], Vincent Andrew Akpan[2, *]

[1]Department of Mathematical Sciences, Kogi State University, Anyigba, Nigeria
[2]Department of Biomedical Technology, The Federal University of Technology, Akure, Nigeria

Abstract

This paper proposes a machine learning approach for information extraction of structured academic data from unstructured web documents. The current challenges of information extraction have been critically examined as well as the state-of-the-art of structured data extraction. The approach used has been simplified and presented using a comprehensive flowchart. The machine learning information extraction scheme was validated using Kogi State University (KSU), Anyigba, Kogi State-Nigeria. The preliminary studies of KSU as well as an organogram of KSU are presented in the paper. The feasibility and realization of the machine learning algorithms for information extraction of structured academic data from unstructured web documents were highlighted and the goals accomplished were also listed.

Keywords

Artificial Neural Networks, Information Extraction, Machine Learning, Structured Academic Data, Unstructured Web Documents

Received: January 4, 2019 / Accepted: September 8, 2019 / Published online: September 21, 2021

1. Introduction

Information extraction (IE) is more about extracting (or inferring) general knowledge (or relations) from a set of documents or information. Note that here all the content of the documents could be considered as a whole corpus of data from where knowledge can be extracted. Of course also for this case somehow possible to specify what is to be extracted, but it is more about properties/relations than specific subjects/topics. Properties are more domain-specific; while generally relations cover more generic scenarios. On the other hand, information retrieval (IR) is about returning the information that is relevant for a specific query or field of interest. Note that this information could also be in the form of general documents, sure enough search engines are a notable example of such task. The most important entities recognizable for information retrieval are the initial set of documents/information and the query that specify "what to search for".

"Information science is the science and practice dealing with the effective collection, storage, retrieval, and use of information. It is concerned with recordable information and knowledge, and the technologies and related services that facilitate their management and use. More specifically, information science is a field of

* Corresponding author
E-mail address: agbogun.jb@ksu.edu.ng (J. B. Agbogun), vaakpan@futa.edu.ng (V. A. Akpan)

professional practice and scientific inquiry addressing the effective communication of information and information objects, particularly knowledge records, among humans in the context of social, organizational, and individual need for and use of information. The domain of information science is the transmission of the universe of human knowledge in recorded form, centering on manipulation (representation, organization, and retrieval) of information, rather than knowing information" [1].

For IE one could instead, for example, ask to extract all the names of cities, or e-mail addresses, that appear in a corpus of documents. It is also possible to go much more generic, asking simply to extract knowledge. It is obvious that this is really generic, but it can be accomplish, for example, by obtaining triplets of the form subject-action-object for each valid sentence of a text (this is best suited for natural language texts). This work is particularly focused on IE.

IE dated back to the late 1970s in the early days of natural language processing (NLP) [2]. An early commercial system from the mid1980s was JASPER, built for Reuters by the Carnegie Group with the aim of providing real-time financial news to financial traders [3]. Beginning in 1987, IE was spurred by a series of Message Understanding Conferences (MUC). The MUC is a competition-based conference that focused on the following domains, namely: *1).* MUC1 (1987) and MUC2 (1989) for naval operations messages; *2).* MUC3 (1991) and MUC4 (1992) for terrorism in Latin American countries; *3).* MUC5 (1993) for joint ventures and microelectronics domain; *4).* MUC6 (1995) for news articles on management changes; and *5).* MUC7 (1998) for satellite launch reports [4]. Considerable support came from the U.S. Defense Advanced Research Projects Agency (DARPA), who wished to automate mundane tasks performed by government analysts, such as scanning newspapers for possible links to terrorism [4].

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction. IE on non-text documents is becoming an increasing topic in research and information extracted from multimedia documents can now be expressed in a high level structure as it is done on text. This naturally led to the fusion of extracted information from multiple kinds of documents and sources. Due to the difficulty of the problem, current approaches to IE focus on narrowly restricted domains.

The world wide web (www) is the world's largest repository of knowledge, and it is being constantly augmented and maintained by millions of people [5, 6]. However, it is in a form intended for human reading, not in a database form with records and fields that can be easily manipulated and understood by computers. In spite of the promise of the Semantic Web, the use of English and other natural language text will continue to be a major medium for communication and knowledge accumulation on the Web, in e-mail, news articles, and elsewhere.

Eventually, a point will be reached at which the answer to almost any question will be available online somewhere, but we will have to wade through more and more material to find it. The next step in improved search tools will be a transition from keyword search on documents to higher-level queries. Queries where the search hits will be objects, such as people or companies instead of simply documents; queries that are structured and return information that has been integrated and synthesized from multiple pages; as well as queries that are stated as natural language questions for example: "*Who were the first three female U.S. Senators?*" and answered with succinct responses [5].

The first half of the Internet revolution consisted of the creation of a wide area network for easy data sharing, enabling human access to an immense store of knowledge and services. The second half of the Internet revolution is yet to come [6]. It will happen when there is machine access to this immense knowledge base, and it becomes possible to perform pattern analysis, knowledge discovery, reasoning, and semi-automated decision-making on top of it [7]. Information extraction will be a key part of the solution making this possible.

Simply, a data is something that provides information about a particular thing and can be used for analysis.

Data can have different sizes and formats. For example, all the information of a particular person in curriculum vitae (CV) or resume including his educational details, personal interests, working experience, address etc. in pdf, docx file format having size in kb's. This is very small-sized data which can be easily retrieved and analyzed. But with the advent of newer technologies in this digital era, there has been a tremendous rise in the data size. Data has grown from kilobytes (KB) to petabytes (PB). This huge amount of data is referred to as big data and requires advance tools and software for processing, analyzing and storing purposes.

The data that has a structure and is well organized either in the form of tables or in some other way and can be easily operated is known as structured data. Searching and accessing information from such type of data is very easy. For example; data stored in relational database in the form of tables having multiple rows and columns. The spreadsheet is a good example of structured data.

The data that has no structure and is unorganized either in the form of tables or some other way and cannot be easily operated is known as unstructured data. Operating on such type of data becomes difficult and requires advance tools and softwares to access information. For example, images and graphics, pdf files, word document, audio, video, emails, powerpoint presentations, web pages and web contents, wikis, streaming data, location coordinates, etc.

Semi-structured data is basically a structured data that is unorganized. Web data such as JSON (JavaScript Object Notation) files, BibTex files, .csv files, tab-delimited text files, XML and other markup languages are the examples of Semi-structured data found on the web. Due to unorganized information, the semi-structured is difficult to retrieve, analyze and store as compared to structured data. It requires software framework like Apache Hadoop to perform all this.

On top of this, there is simply much more unstructured data than structured. Unstructured data makes up 80% and more of enterprise data, and is growing at the rate of 55% and 65% per year. And without the tools to analyze this massive data, organizations are leaving vast amounts of valuable data on the business intelligence table as illustrated in Table 1 [8, 9].

Structured data is far easier for Big Data programs to digest, while the myriad formats of unstructured data create greater challenges. Yet both types of data play a key role in effective data analysis. Again, structured data is traditionally easier for Big Data applications to digest and yet today's data analytics solutions are making great strides in this area.

The neural network-based machine learning (NN-ML) techniques is used because it is the application of artificial intelligence whereby available information is used through algorithms to process data and the machine learning algorithms have strong mathematical and statistical basis that do not use to take domain knowledge of data and pre-processing into account.

Table 1. Examples of structured and unstructured data and the possible sources of their generation.

S/N	Sources	Structured Data	Unstructured Data
1.	Characteristics	1. Pre-defined data models 2. Usually text only 3. Easy to search	1. No pre-defined data model 2. May be text, images, sound, video or other formats 3. Difficult to search
2.	Resides in	1. Relational databases 2. Data warehouses	1. Applications 2. No SQL databases 3. Data warehouses 4. Data lakes
3.	Generated by	1. Human or machines	1. Human or machines
4.	Typical applications	1. Airline reservation systems 2. Inventory control 3. CRM systems 4. ERP systems	1. Word processing 2. Presentation software 3. Email clients 4. Tools for viewing or editing media
5.	Examples	1. Dates 2. Phone numbers 3. Social security numbers 4. Credit card numbers 5. Customer names 6. Addresses 7. Product names and numbers 8. Transaction information	1. Text files 2. Reports 3. Email messages 4. Audio files 5. Video files 6. Images 7. Surveillance images

Based on the leaning that happened in the previous stages, from the input dataset fed to the system, predications are made. As such machine learning is used in order to make very good predictions that are good enough to be useful. Neural Network-based Supervised Machine Learning Algorithm based on The Teacher-Forcing method was employed.

2. Literature Review

2.1. Preliminary Studies

The main task of applying information extraction on text is linked to the problem of text simplification in order to create a structured view of the information present in free text. The overall goal is to create a more easily machine-readable text to process these sentences. Typical sub-tasks of IE include [4, 9]:

1). Named entity extraction which could include:

*i).*Named entity recognition: recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions, employing existing knowledge of the domain or information extracted from other sentences. Typically the recognition task involves assigning a unique identifier to the extracted entity. A simpler task is *named entity detection*, which aims to detect entities without having any existing knowledge about the entity instances. For example, in processing the sentence "M. Smith likes fishing", *named entity detection* would denote detecting that the phrase "M. Smith" does refer to a person, but without necessarily having (or using) any knowledge about a certain *M. Smith* who is (or, "might be") the specific person whom that sentence is talking about.

ii). Coreference resolution: detection of coreference and anaphoric links between text entities. In IE tasks, this is typically restricted to finding links between previously-extracted named entities. For example, "International Business Machines (IBM)" and "IBM" refer to the same real-world entity. If the two sentences are taken "M. Smith likes fishing. But he doesn't like biking", it would be beneficial to detect that "he" is referring to the previously detected person "M. Smith".

iii). Relationship extraction: identification of relations between entities, such as:

a). PERSON works for ORGANIZATION (extracted from the sentence "Bill works for IBM").

b). PERSON located in LOCATION (extracted from the sentence "Bill is in France").

2). Semi-structured information extraction which may refer to any IE that tries to restore some kind information structure that has been lost through publication such as:

*i).*Table extraction: finding and extracting tables from documents; and

ii). Comments extraction: extracting comments from actual content of article in order to restore the link between author of each sentence

3). Language and vocabulary analysis

i). Terminology extraction: finding the relevant terms for a given corpus

4). Audio extraction

Table 2. Sample extraction rule learned by RAPIER.

S/N	Pre-Filter Pattern	Pre-Filter Pattern	Pre-Filter Pattern
1.	syntactic: {nn.nnp}	word: undisclosed	semantic: price
2.	list: length 2	syntactic: jj	

i). Template-based music extraction: finding relevant characteristic in an audio signal taken from a given repertoire. For instance in [10], time indexes of occurrences of percussive sounds can be extracted in order to represent the essential rhythmic component of a music piece.

Note that many approaches combine multiple subtasks of IE in order to achieve a wider goal. Machine learning, statistical analysis and/or natural language processing are often used in IE.

IE on non-text documents is becoming an increasing topic in research and information extracted from multimedia documents can now be expressed in a high level structure as it is done on text. This naturally led to the fusion of extracted information from multiple kinds of documents and sources.

Furthermore, three standard approaches are now widely accepted for IE from structured and non-structured text as well as IE from web pages. These approaches include:

1). Hand-written regular expressions (perhaps stacked)

2). Using classifiers:

i). Generative: naïve Bayes classifier; and

ii). Discriminative: maximum entropy models such as Multinomial logistic regression.

3). Sequence models

*i).*Hidden Markov model;

*ii).*Conditional Markov model (CMM) / Maximum-entropy Markov model (MEMM); and

iii). Conditional random fields (CRF) are commonly used in conjunction with IE for tasks as varied as extracting information from research papers to extracting navigation instructions [11, 12].

It should be noted that several other numerous other approaches exist for IE including hybrid approaches that combine some of the standard approaches previously listed above in this sub-section.

2.2. Information Extraction (IE) Methods

There are a variety of approaches to constructing IE systems. One approach is to manually develop information-extraction rules by encoding patterns (e.g. regular expressions) that reliably identify the desired entities or relations. For example, the Suiseki system extracts information on interacting proteins from biomedical text using manually developed patterns [13].

However, due to the variety of forms and contexts in which the desired information can appear, manually developing patterns is very difficult and tedious and rarely results in robust systems. Consequently, supervised machine-learning method trained on human annotated corpora has become the most successful approach to developing robust IE systems [14]. A variety of learning methods have been applied to IE.

One approach is to automatically learn pattern-based extraction rules for identifying each type of entity or relation. For example, the previously developed system by Rapier learns extraction rules consisting of three parts [15, 16]: *1).* A pre-filler pattern that matches the text immediately preceding the phrase to be extracted, *2).* a filler pattern that matches the phrase to be extracted, and *3).* a post-filler pattern that matches the text immediately following the filler. Patterns are expressed in an enhanced regular-expression language, similar to that used in Perl and a bottom-up relational rule learner is used to induce rules from a corpus of labeled training examples [17]. In Wrapper Induction [18] and Boosted Wrapper Induction (BWI) [19], regular-expression-type patterns are learned for identifying the beginning and ending of extracted phrases. Inductive Logic Programming (ILP) has also been used to learn logical rules for identifying phrases to be extracted from a document [20, 21].

An alternative general approach to IE is to treat it as a sequence labeling task in which each word (token) in the document is assigned a label (tag) from a fixed set of alternatives. For example, for each slot, X, to be extracted, a token label is included BeginX to mark the beginning of a filler for X and Inside X to mark other tokens in a filler for X. Finally, the label order is included for tokens that are not included in the filler of any slot. Given a sequence labeled with these tags, it is easy to extract the desired fillers.

One approach to the resulting sequence labeling problem is to use a statistical sequence model such as a Hidden Markov Model (HMM) [22] or a Conditional Random Field (CFR) [23]. Several earlier IE systems used generative HMM models; however, discriminately-trained CRF models have recently been shown to have an advantage over HMM's [24–27]. In both cases, the model parameters are learned from a supervised training corpus and then an efficient dynamic programming method based on the Viterbi algorithm is used to determine the most probable tagging of a complete test document [28].

Another approach to the sequence labeling problem for IE is to use a standard feature-based inductive classifier to predict the label of each token based on both the token itself and its surrounding context. Typically, the context is represented by a set of features that include the one or two tokens on either side of the target token as well as the labels of the one or two preceding tokens (which will already have been classified when labeling a sequence from left to right). Using this general

approach, IE systems have been developed that use many different trained classifiers such as decision trees [29], boosting [30], memory-based learning (MBL) [31], support-vector machines (SVMs) [32], maximum entropy (MaxEnt) [33], transformation-based learning (TBL) [34] and many others [35].

Many IE systems simply treat text as a sequence of uninterpreted tokens; however, many others use a variety of other NLP tools or knowledge bases. For example, a number of systems preprocess the text with a part-of-speech (POS) tagger (e.g. [36, 37]) and use words' POS (e.g. noun, verb, adjective) as an extra feature that can be used in handwritten patterns [13], learned extraction rules [16], or induced classifiers [35]. Several IE systems use phrase chunkers to identify potential phrases to extract [35, 38, 39]. Others used complete syntactic parsers, particularly those which try to extract relations between entities by examining the syntactic relationship between the phrases describing the relevant entities [40–42]. Some use lexical semantic databases, such as WordNet [43], which provide word classes that, can be used to define more general extraction patterns [16].

As a sample extraction pattern, Table 2 shows a rule learned by Rapier for extracting the transaction amount from a newswire concerning a corporate acquisition [16]. This rule extracts the value "undisclosed" from phrases such as "sold to the bank for an undisclosed amount" or "paid Honeywell an undisclosed price". The pre-filler pattern matches a noun or proper noun (indicated by the POS tags 'nn' and 'pn', respectively) followed by at most two other unconstrained words. The filler pattern matches the word "undisclosed" only when its POS tag is "adjective". The post-filler pattern matches any word in WordNet's semantic class named "price".

2.3. State-of-the-Art in Information Extraction

In 2003, Arasu and Garcia-Molina from Stanford University wrote on Extracting structured data from web pages [6]. They said the earliest information extraction techniques rely on a human to encode knowledge of the template into a program called wrapper. Their goal was to deduce the template without any human input and use the deduced template to extract data. The paper presented an algorithm called EXALG for extracting structured data from a collection of web pages generated from a common template. EXALG first discovers the unknown template that generated the pages and uses the discovered template to extract the data from the input pages. EXALG uses two novel concepts, equivalence classes and differentiating roles, to discover the template. It has been shown in [6] that experiments on several collections of web pages, drawn from many well-known data rich sites indicated that EXALG is extremely good in extracting the data from the web pages. Another desirable feature of EXALG is that it does not completely fail to extract any data even when some of the assumptions made by EXALG are not met by the input collection. In other words the impact of the failed assumptions is limited to a few attributes.

Andrew McCallum stated that the U.S. Department of Labor course extraction problem was solved by a company called WhizBang Labs using a combination of several machine-learning components [44]. To find the Web pages likely to contain course listings, text classification was used in conjunction with a spider. Statistical language modeling methods hypothesized segmentations and classifications of the different fields, which also were put into a classifier responsible for coarse-scale segmentation of one course from another. A method called scoped learning was then used to learn formatting (wrapper-like) regularities on the fly from each page, without human intervention. Logistic-regression classifiers were used to complete the association and deduplication phases. (Conditional random fields were not used only because they had not yet been developed.) In the end, the project was deemed a success–data was extracted with sufficient accuracy so that it could be deposited directly into the Web site's structured database.

According to Tang and co-workers information is hidden in the large volume of web pages and thus it is necessary to extract useful information from the web content, called Information Extraction [45]. In information extraction, given a sequence of instances, they identified and pull out a sub-sequence of the input that represents information they were interested in. In the past years, there was a rapid expansion of activities in the information extraction area. Many methods have been proposed for automating the process of extraction.

However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still presents many challenging research problems. They investigated the problems of information extraction and surveyed existing methodologies for solving these problems. Several real-world applications of information extraction were introduced. Emerging challenges were also discussed.

On 21st April 2014, the Visual Web Ripper V2.123.0 was released while on 23rd April 2014, the Visual Web Ripper V2.123.2 was also released [46]. Visual Web Ripper is a powerful visual tool used for automated web scraping, web harvesting and content extraction from the web. This data extraction software can automatically walk through whole web sites and collect complete content structures such as product catalogues or search results [46].

Sunandan and co-workers, presented a solution to the problem of structuring unstructured online ads with a < key; value > style representation [47]. They proposed a graph-based unsupervised algorithm which gave a performance with an accuracy of 67.74% for cars and 68.74% for apartment ads downloaded from Craigslist. They also presented an alternative supervised learning algorithm where they used conditional random field (CRF) to compute the most probable label sequence given an observed sequence of words in an ad. The supervised algorithm achieved an accuracy of 74.07% and 72.59% respectively for car and apartment ads. Lower accuracies in the unsupervised method can be attributed to the fact that there are some aspects to the problem which are very difficult to model in an unsupervised method. Implementation of the supervised algorithm actually shows that some of the shortcomings of the unsupervised method can be reduced by the supervised method. They are currently exploring the possibilities to further enhance the accuracy of their algorithms.

In the work of Gowri and his team they wrote that Text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources [48]. In the current scenario, text classification gains lot of significance in processing and retrieval of text. Automated document classification becomes a key technology to deal and organize huge volume of documents and it frees organizations from the need of manually organizing document bases. A traditional approach to text categorization requires encoding documents into numerical vectors. This type of traditional document encoding causes two main problems: huge dimensionality and sparse distribution. Information retrieval techniques such as text indexing have been developed to handle the unstructured documents. The related task information extraction (IE) is about specific items in natural language documents.

Arvinder and Deepti noted that most of the data is in the form of text these days [49]. While databases store only structured data, most of the data is unstructured like text documents, web pages, emails etc. Text mining is what is required if useful information needs to be extracted from tons of text. But where to begin, what are the popular tools, which techniques are used, what are the features. Beginning is always the toughest, so their paper tries to explore the tools available for text mining to help new researchers and practitioners in the field of text mining.

The work on information extraction in illicit web domains showed that extracting useful entities and attribute values from illicit domains such as human trafficking is a challenging problem with the potential for widespread social impact [50]. Such domains employ a typical language models, have `long tails' and suffer from the problem of concept drift. In their paper, they proposed a lightweight, feature-agnostic Information Extraction (IE) paradigm specifically designed for such domains. Their approach uses raw, unlabeled text from an initial corpus, and a few (12-120) seed annotations per domain-specific attribute, to learn robust IE models for unobserved pages and websites. Empirically, they demonstrated that their approach can outperform feature-centric Conditional Random Field baselines by over 18% F-Measure on five annotated sets of real-world human trafficking datasets in both low-supervision and high-supervision settings. They also showed that their approach is demonstrably robust to concept drift, and can be efficiently bootstrapped even in a serial computing environment.

Furthermore, Yanshan and co-workers pointed out that with the rapid adoption of electronic health records

(EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research [51]. One critical component used to facilitate the secondary use of EHR data is the information extraction (IE) task, which automatically extracts and encodes clinical information from text. Their Objectives: In their literature review, they presented a review of recent published research on clinical information extraction (IE) applications. The authors conducted a literature search for articles published from January 2009 to September 2016 based on Ovid MEDLINE In-Process as well as Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and ACM Digital Library. Finally, it has been shown in [51] that a total of 1917 publications were identified for title and abstract screening. Of these publications, 263 articles were selected and discussed in their review in terms of publication venues and data sources, clinical IE tools, methods, and applications in the areas of disease- and drug-related studies, and clinical workflow optimizations. Their conclusion was that clinical IE has been used for a wide range of applications; however, there is a considerable gap between clinical studies using EHR data and studies using clinical IE. Their study enabled them to gain a more concrete understanding of the gap and to provide potential solutions to bridge this gap.

3. The Architecture of the Scheme for Unstructured Information Extraction

3.1. Open Information Extraction

Information-extraction (IE) systems seek to distil semantic relations from natural-language text, but most systems use supervised learning of relation-specific examples and are thus limited by the availability of training data. *Open* IE systems such as TextRunner, on the other hand, aim to handle the unbounded number of relations found on the Web [52–54]. But how well can these open systems perform?

An open information extractor is a function from a document, d, to a set of triples, $\{\langle \text{arg}_1, rel, \text{arg}_2 \rangle\}$, where the args are noun phrases and rel is a textual fragment indicating an implicit, semantic relation between the two noun phrases. The extractor should produce one triple for every relation stated *explicitly* in the text, but is not required to infer implicit facts. It has been assumed that all relational instances are stated within a single sentence [55]. Note the difference between open IE and the traditional approaches (e.g., as in WebKB), where the task is to decide whether some pre-defined relation holds between (two) arguments in the sentence [53, 54].

Furthermore, Wu and Weld made an attempt to learn an open extractor *without direct supervision*, i.e. without annotated training examples or hand-crafted patterns [55]. Their input was Wikipedia, a collaboratively-constructed encyclopaedia. As output, Wikipedia-based Open Extractor (WOE) produces an unlexicalized and relation-independent open extractor. Their objective was to construct an extractor which generalizes beyond Wikipedia, handling other corpora such as the general Web.

3.2. Structure of Kogi State University

Before the university website was designed, Figure 1 is the organogram of Kogi State University. This is the basic structure of KSU to show you the schematic of how KSU looks like. Therefore with the advent of internet and the need to develop a website for Kogi State University, based on the organogram in Figure 1, the website was developed. Figure 2 is the same organogram but removing the arrow branches to other links showing only the links to the area of interest to this research work.

The system was implemented using the Structure of Kogi State University website (Figure 1) which was reduced to the area of case study (Figure 2). Furthermore, Figure 1 shows the information of the complete Structure of Kogi State University Website in the form of an organogram, including programmes, students and Scholars (staff). Figure 2 contains some WebPages of Kogi State University (areas of case study) corresponding to the structure in Figure 1, with other links removed but contains only areas of case study WebPages of Kogi State University corresponding to the structure in Figure 2 with all other unwanted links removed. The task is to extract important data of Figure 1 from the data source of webs

corresponding to Figure 2 and to organize the data with relationships of reality. In the university, the most important units are: the Scholars (staff), students including the Programmes, department and laboratory, the most important people are the staff members and they work in different faculties and departments with different identities.

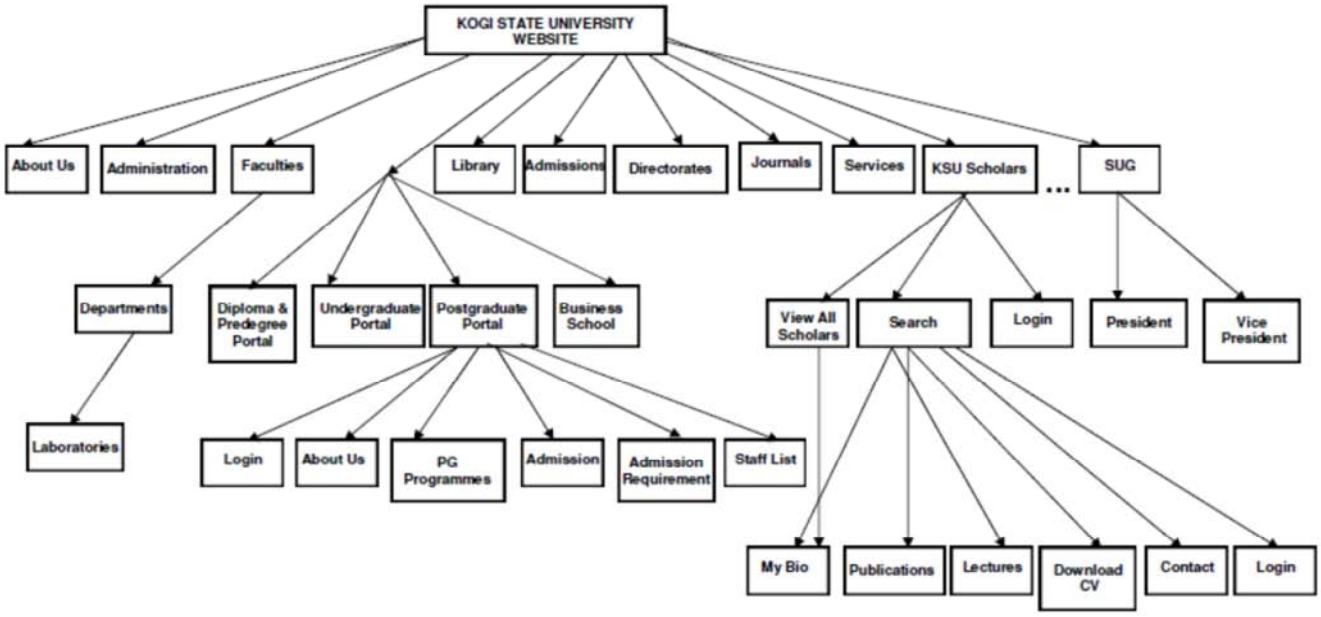

Figure 1. Structure of Kogi State University website.

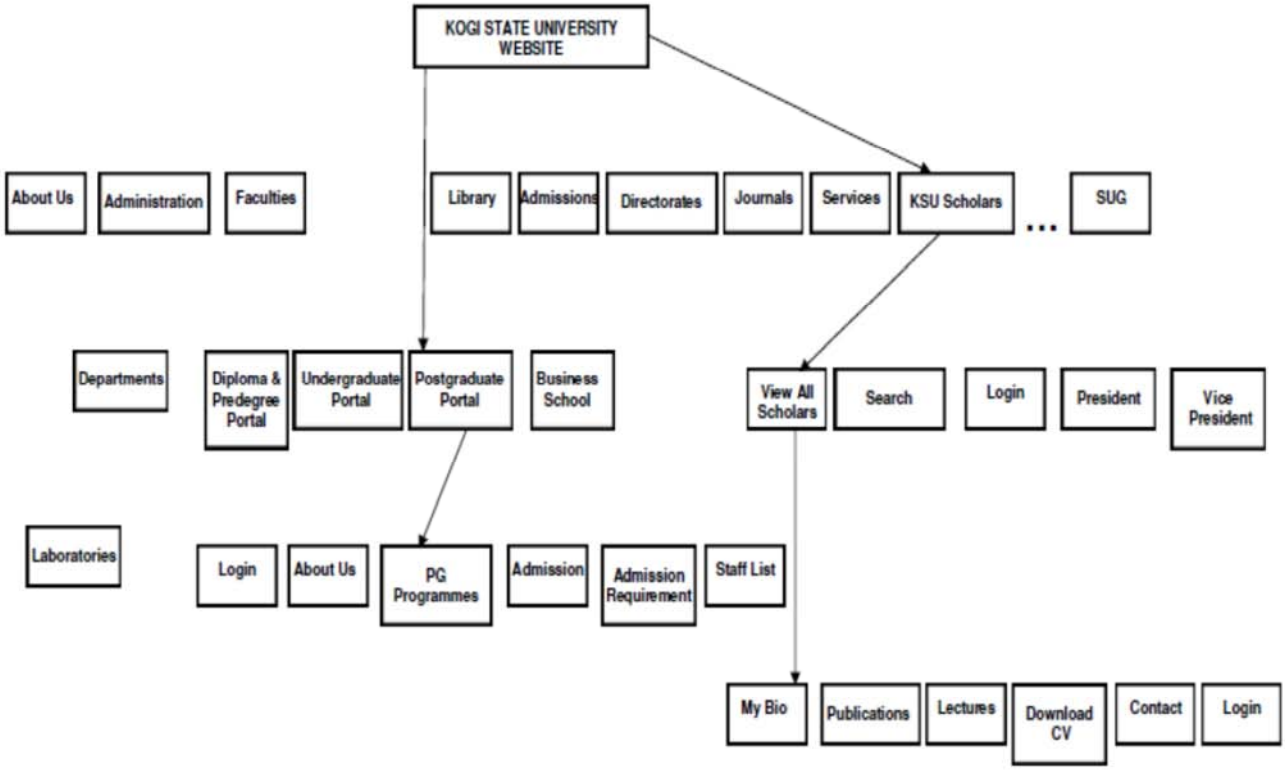

Figure 2. Structure of KSU site showing links to the case study area only.

KSU Scholars block in Figure 2 is one of the web pages on KSU website. It contains the following options: Home, View All Scholars, Login, Contact, Search any keyword and Search Name(s) of Scholar. It is from this webpage that View All Scholars can be reached. The link is www.scholars.ksu.edu.ng

3.3. Instances of KSU Website

1) KSU Diploma & Predegree Portal (CPDS)

The main page of the KSU Diploma & Predegree programme portal is shown in Figure 3 which has the following menu options: Home, About Us, Programmes, Admission, Transcript, Student Login, Help desk, Portal Home, Screening Officer, Official website. It also displays further information like welcome to CPDS official website, vision and mission statement and general information from KSU News centre. All the menu options listed and all other links can be accessed from here. The web link is http://cpds.ksu.edu.ng/

Figure 3. KSU Diploma and Predegree Portal (CPDS).

2) KSU Undergraduate Portal (100 to 400 Level Students)

The main page of the Undergraduate portal for KSU is displayed in Figure 4. It has the following menu options: Home, Instruction, Calendar, Student E-Mail, iTranscript, Help desk, FAQ, Student Login, Registration Steps, Help Desk, Official website, Portal Home, Screening Officer. It also displays further information like Quick Links and Hot News. All the menu options listed and all other links can be accessed from here. The web link is http://portal.ksu.edu.ng/.

3) KSU Business School Portal

Figure 5 is the main page of the KSU Business School Portal. It has the following menu options: Home, About Us, Programmes, Admission, Admission requirement, Our staff, Alumni, Tuition Fee, Contact Us, Official website, KSU Scholars, Student Login. It also displays further information like welcome to Kogi State University Business School and its history and general information from KSU News centre. All the menu options listed on that page and all other links can be accessed from here. The web link is http://kbs.ksu.edu.ng/.

3.4. The Implementations Flowchart

This work aims at the development of neural network-based machine learning algorithms for information extraction of structured academic data from unstructured web document(s) and the flowchart for the methodology to achieve the aim is as shown in Figure 6.

Figure 4. KSU Undergraduate Portal (100 to 400 level students).

Figure 5. KSU Business school portal.

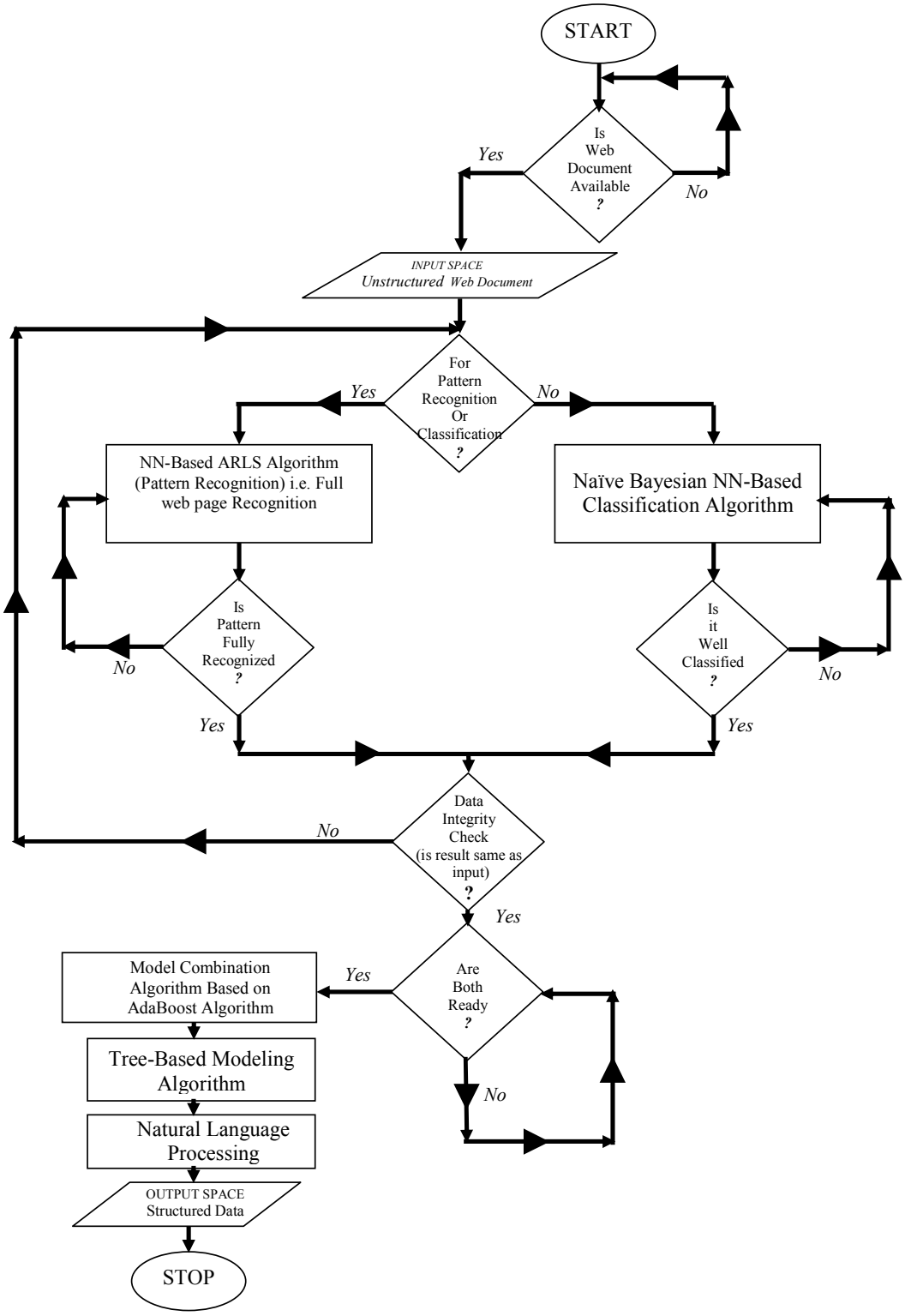

Figure 6. Flowchart of the machine learning information extraction scheme.

The application will wait for the webpage by checking if the webpage is available. If the webpage is not available, it will continue to check until the webpage is available. When available, it will show the webpage and will accept it as INPUT and continue to the next stage. At the next stage the full web page will be recognized by passing through a Neural Network-Based Adaptive Recursive Least Squares (ARLS) Algorithm based on Teacher-Forcing method in order to recognize the pattern, at the same time, the full webpage will be classified by passing through the Naïve Bayesian Neural Network-Based Classification Algorithm for

pattern classification based on posteriori parameter distribution with hyper-parameter optimization. Data integrity check was performed on the two results to be sure that there is no lost of data (i.e. is the result Dataset same as the input Dataset?) after which the two results were combined with a model combination algorithm based on AdaBoost Algorithm after which the result of the combination passes through a Binary Tree-Based Model for partitioning the input space (i.e. unstructured data) to the desired output space (i.e. structured data).The machine learning Algorithms for information extraction of structured academic data from unstructured web document(s) was implemented and deployed by using Natural Language Processing (NLP) Scheme.

4. Discussions

While information extraction has applications in a wide range of domains, the specific type and structure of the information to be extracted depend on the need of the particular application(s) [56]. Some example applications of information extraction include but not limited to the following:

1) Biomedical researchers often need to sift through a large amount of scientific publications to look for discoveries related to particular genes, proteins or other biomedical entities. To assist this effort, simple search based on keyword matching may not suffice because biomedical entities often have synonyms and ambiguous names, making it hard to accurately retrieve relevant documents. A critical task in biomedical literature mining is therefore to automatically identify mentions of biomedical entities from text and to link them to their corresponding entries in existing knowledge bases such as the FlyBase;
2) Financial professionals often need to seek specific pieces of information from news articles to help their day-to-day decision making. For example, a finance company may need to know all the company takeovers that take place during a certain time span and the details of each acquisition. Automatically finding such information from text requires standard information extraction technologies such as named
3) Intelligence analysts review large amounts of text to search for information such as people involved in terrorism events, the weapons used and the targets of the attacks. While information retrieval technologies can be used to quickly locate documents that describe terrorism events, information extraction technologies are needed to further pinpoint the specific information units within these documents; and
4) With the fast growth of the Web, search engines have become an integral part of people's daily lives, and users' search behaviours are much better understood now. Search based on bag-of-word representation of documents can no longer provide satisfactory results. More advanced search problems such as entity search, structured search and question answering can provide users with better search experience. To facilitate these search capabilities, information extraction is often needed as a pre-processing step to enrich document representation or to populate an underlying database.
5) The problem studied in this paper also has certain resemblances to the works of Kejriwal and Szekely [50] and Dong and co-workers [57] but a different approach will be proposed study. To the best of our knowledge, the combined use of neural network-based supervised machine learning algorithm based on the teacher-forcing method for pattern recognition with the Naïve Bayesian classification algorithm using Bayesian model averaging technique with AdaBoosting Algorithm has not been studied in prior work.

5. Conclusion and Recommendation

The main aim of this research was focused on the development of machine learning algorithms for information extraction of structured academic data from unstructured web documents which were accomplished according to but not necessarily limited to the following goals:

1) a robust neural network-based supervised machine learning algorithm based on the teacher-forcing method for pattern recognition was formulated;
2) a Naïve Bayesian neural network classification algorithm for pattern classification based on

posteriori parameter distribution with hyper-parameter optimization was also formulated;

3) the neural network-based supervised machine learning algorithm based on teacher-forcing pattern recognition was combined with the Naïve Bayesian classification algorithm using Bayesian model Averaging technique with AdaBoosting Algorithm;

4) a binary Tree-based model for partitioning the input space to the desired output space was developed;

5) the machine learning algorithms for information extraction of structured academic data from unstructured web documents was implemented and deployed using natural language processing (NLP) scheme; and

6) the performances of the machine learning algorithms were compared with a standard automatic information extraction algorithm for given document(s).

References

[1] T. Saracevic, (2009). Information Science. In M. J. Bates (ED.), Encyclopaedia of library and information sciences (3rd ed.) (pp. 2570-2585). New York: Taylor and francis.

[2] P. M. Andersen, P. J. Hayes, A. K. Huettner, L. M. Schmandt, I. B. Nirenburg andS. P. Weinstein,(1992): "Automatic Extraction of Facts from Press Releases to Generate News Stories", 1992 ANLC'92 Proceedings of the third conference of the Applied Natural Language Processing, 1992, pp. 170-177.

[3] J. Cowie and Y. Wilks, "Information Extraction", Retrieved on Saturday 13th January, 2018 from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.6480., 1996.

[4] Wiki, "Information Extraction", Wikipedia-The Free Encyclopedia, Retrieved 6th April, 2018, Available [Online]: https://en.wikipedia.org/wiki/Information_extraction, 2018.

[5] A. Akbik and J. Broß. "Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns", In Proceedings of WWW Workshop, 2009. pp. 205-216.

[6] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages" Proceedings of the 2003 Association for Computing Machinery Special Interest Group on Management of Data (ACM SIGMOD),International Conference on Management of Data, 2003. pp 337-348.

[7] T. Berners-Lee, "TED: Talk on the Next Web". Retrieved on Saturday 13th January, 2018 from: https://www.ted.com/talks/tim_berners_lee_on_the_next_web, 2009.

[8] C. C. Aggarwal and C. X. Zhai, "Mining Text Data",DOI 10.1007/978-1-4614-3223-4_2, © Springer Science+Business Media, LLC 2012.

[9] D. Freitag, "Machine Learning for Information Extraction in Informal Domains", Kluwer Academic Publishers. Printed in The Netherlands, 2000.

[10] A. Zils, F. Pachet, O. Delerue and F. Gouyon, "Automatic Extraction of Drum Tracks from Polyphonic Music Signals (http://www.csl.sony.fr/downloads/papers/2002/ZilsMusic.pdf), In Proceedings of WedelMusic, Darmstadt, Germany, 2002.

[11] F. Peng and A. McCallum, A. (2006): "Information extraction from research papers using conditional random fields", Information Processing & Management, vol. 42, no. 4, 2006, pp. 963. doi: 10.1016/j.ipm.2005.09.002.

[12] N. Shimizu and A. Hass, "Extracting Frame-based Knowledge Representation from Route Instructions", 2006. Retrieved on Saturday 13th January, 2018 from https://pdfs.semanticscholar.org/fb72/c577ef096d9705ba26e21be0a3db93c6500b.pdf

[13] C. Blaschke and A. Valencia, "The frame-based module of the Suiseki information extraction system", IEEE Intelligent Systems, vol. 17, 2002, pp. 14-20.

[14] C. Cardie, "Empirical methods in information extraction", AI Magazine, vol. 18, no. 4, 1997, pp. 65-79.

[15] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), Orlando, FL, July 1999, pp. 328-334.

[16] M. E. Califf and R. J. Mooney, "Bottom-up relational learning of pattern matching rules for information extraction", Journal of Machine Learning Research, 4: 177-210, 2003.

[17] L. Wall, T. Christiansen and R. L. Schwartz, "Programming Perl", O'Reilly and Associates, Sebastopol, CA, 1996.

[18] N. Kushmerick, D. S. Weld and R. B. Doorenbos, "Wrapper induction for information extraction", In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, 1997, pp. 729-735.

[19] D. Freitag and N. Kushmerick, "Boosted wrapper induction", In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, July 2000. AAAI Press/The MIT Press, pp. 577-583.

[20] S. H. Muggleton, "Inductive Logic Programming", Academic Press, New York, NY, 1992.

[21] D. Freitag, "Toward general-purpose learning for information extraction", In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and COLING-98 (ACL/COLING-98), Montreal, Quebec, 1998, pp. 404-408.

[22] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", In Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[23] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", In Proceedings of 18th International Conference on Machine Learning (ICML-2001), Williamstown, MA, 2001, pp. 282-289.

[24] D. M. Bikel, R. Schwartz, and R. M. Weischedel, "An algorithm that learns what's in a name", Machine Learning, 34: 211-232, 1999.

[25] D. Freitag and A. McCallum, "Information extraction with HMM structures learned by stochastic optimization", In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, AAAI Press/The MIT Press, 2000.

[26] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields", In Proceedings of Human Language Technology Conference/North American Association for Computational Linguistics Annual Meeting (HLT-NAACL-2004), Boston, MA, 2004.

[27] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction", In Advances in Neural Information Processing Systems 17, Vancouver, Canada, 2005.

[28] A. J. Viterbi, "Error bounds for convolutional codes and and asymptotically optimum decoding algorithm", IEEE Transactions on Information Theory, vol. 13, no. 2, pp. 260-269, 1967.

[29] S. W. Bennett, C. Aone, and C. Lovell, "Learning to tag multilingual texts through observation, In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97),, Providence, RI, 1997, pp. 109-116.

[30] X. Carreras, L. M`arquez, and L. Padr´o, "A simple named entity extractor using AdaBoost", In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.

[31] F. D. Meulder and W. Daelemans, "Memory-based named entity recognition using unannotated data", In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.

[32] J. Mayfield, P. McNamee, and C. Piatko, "Named entity recognition using hundreds of thousands of features", In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.

[33] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum entropy approach", In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), pages 160-163, Edmonton, Canada, 2003.

[34] L. Tanabe and W. J. Wilbur, "Tagging gene and protein names in biomedical text", Bioinformatics, vol. 18, no. 8, pp. 1124-1132, 2002.

[35] E. F. T. K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition", In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.

[36] K. W. Church, "A stochastic parts program and noun phrase parser for unrestricted text", In Proceedings of the Second Conference on Applied Natural Language Processing, Austin, TX, Association for Computational Linguistics, 1988, pp. 136-143.

[37] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging", Computational Linguistics, vol. 21, no. 4, pp. 543-565, 1995.

[38] D. Zelenko, C. Aone, and A. Richardella, "Kernel method for relation extraction", Journal of Machine Learning Research, vol. 3, 2003, pp. 1083-1106.

[39] L. A. Ramshaw and Mitch P. Marcus, "Text chunking using transformation-based learning", In Proceedings of the 3rd Workshop on Very Large Corpora, 1995, pp. 82-94.

[40] M. J. Collins, "Three generative, lexicalised models for statistical parsing", In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), 1997, pp. 16-23.

[41] A. Culotta and J. Sorensen, "Dependency tree kernels for relation extraction", In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, July 2004.

[42] S. Ray and M. Craven. "Representing sentence structure in hidden Markov models for information extraction", In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA, 2001, pp. 1273-1279.

[43] C. D. Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, Cambridge, MA, 1998.

[44] A. McCallum, A. (2005): "Information Extraction: Distilling structured data from unstructured text in a Magazine", Queue-Social Computing, ACM New York, NY, USA, vol. 3, no. 9, November 2005, pp. 48-57.

[45] J. Tang, M. Hong, D. Zhang, B. Liang and J. Li, "Information Extraction: Methodologies and Applications", In the book of Emerging Technologies of Text Mining: Techniques and Applications, Hercules A. Prado and Edilson Ferneda (Ed.), Idea Group Inc., Hershey, USA, 2007, pp. 1-33. http://keg.cs.tsinghua.edu.cn/jietang/publications/Tang-et-al-Information_Extraction.pdf

[46] Sequentum, Visual Web Ripper V2.123.2 (released on 23rd of April 2014,) downloaded from"www.visualwebripper.com". http://www.sequentum.com/-See more at: http://www.windows8downloads.com/win8-visual-web-ripper-zmsizlqf/#sthash.GBimouoO.dpuf

[47] C. Sunandan, S. Lakshminarayanan and N. Yaw, "Extraction of (Key,Value) Pairs from Unstructured Ads.", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2014 Retreived from https://www.aaai.org/ocs/index.php/FSS/FSS14/paper/viewFile/9196/9080

[48] S. C. Gowri, Dr. K.M. Sundaram (2015), "A Study on Information Retrieval and Extraction for Text Data Words using Data Mining Classifier", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4 no. 10, October 2015, pp. 121-126.

[49] K. Arvinder and C. Deepti, "Comparison of Text Mining Tools", In Proceedings of the 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), AIIT, Amity University Uttar Pradesh, Noida, India, Sep. 7-9, 2016.

[50] M. Kejriwal and P. Szekely, "Information Extraction in Illicit Web Domains", International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License, WWW'17 Perth, Australia. ACM, 2017 978-1-4503-4913-0/17/04, http://dx.doi.org/10.1145/3038912.3052642

[51] W. Yanshan, W. Liwei,R. M. Majid, M. Sungrim, S. Feichen, A. Naveed, L. Sijia, Z. Yuqun, M. Saeed M, S. Sunghwan and L. Hongfang, "Clinical Information Extraction Applications: A literature review", Journal of Biomedical Informatics, vol. 77, 2018, pp. 34-49. https://doi.org/10.1016/j.jbi.2017.11.011

[52] A. McCallum and D. Jensen, "A note on the unification of information extraction and data mining using conditional-probability, relational models", In Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data, Acapulco, Mexico, Aug. 2003.

[53] S. Abteboul, P. Buneman and P. Suciu, "Data on the Web: From Relations to Semi-Structured Data and XML", The Morgan Kaufmann Series in Data Management, 26th October, 1999.

[54] A. L. Bergert, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing", Computational Linguistics, vol. 22, no. 1, pp. 39-71, March 1996.

[55] F. Wu and D. S. Weld, "Open information extraction using Wikipedia", In Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, Uppsala, Sweden, 11th-16th July, 2010, pp. 118-127.

[56] S. Brin, "Extracting patterns and relations from the World Wide Web", In Proceedings of the 1998 International Workshop on the Web and Databases, 1998.

[57] F. Dong, M. Liu and Y. Li (2013), "Automatic Extraction of Semi-structured Web Data", International Journal of Database Theory and Application. Vol. 6, No. 4, pp. 131-144, August, 2013.

Biography

Joshua Babatunde Agbogun received his M.Sc. degree in Computer Science from University of Nigeria Nsukka, Nigeria in 2010. He is currently a Lecturer I with the Department of Mathematical Sciences, Kogi State University, Anyigba, Nigeria and currently working towards his Ph.D at Kogi State University, Anyigba. His research interest is in Machine Learning. He is the author of a book and has authored and/or co-authored 11 articles in refereed Journals and conference proceedings. Mr. Agbogun is a member of Nigeria Computer Society (NCS).

Vincent Andrew Akpan holds a Ph.D. degree in Electrical & Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece in 2011. He is currently a Senior Lecturer and the Head of Biomedical Technology Department, FUTA, Akure, Nigeria. His research interest is in computational intelligence. He is the co-author of a book and has authored and/or co-authored more than 70 articles in refereed journals and conference proceedings. Dr. Akpan is a member of many local and international professional bodies.