Domain Independent Natural Language Processing – A Case Study for Hospital Readmission with COPD

Ankur Agarwal, Ph.D.
Florida Atlantic University
Boca Raton, Florida
aagarwa2@fau.edu

Ravi S Behara, Ph.D.
Florida Atlantic University
Boca Raton, Florida
rbehara@fau.edu

Sirish Malpura
Florida Atlantic University
Boca Raton, Florida
vtyagi@fau.edu

Vivek Tyagi
Florida Atlantic University
Boca Raton, Florida
vtyagi@fau.edu

Abstract—**Natural language processing is a field of computer science, which focuses on interactions between computers and human (natural) languages. The human languages are ambiguous unlike Computer languages, which make its analysis and processing difficult. Most of the data present these days is in unstructured form (such as: Accident reports, Patient discharge summary, Criminal records etc), which makes it hard for computers to understand for further use and analysis. This unstructured text needs to be converted into structured form by clearly defining the sentence boundaries, word boundaries and context dependent character boundaries for further analysis. This paper proposes a component-based domain-independent text analysis system for processing of the natural language known as Domain-independent Natural Language Processing System (DINLP). Further the paper discusses the system capability and its application in the area of bioinformatics through the case study**

Keywords— Chronic Obstructive Pulmonary Disease; Natural Language Processing; Term Extraction

I. INTRODUCTION

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid.

Symbolic Approach: Symbolic approaches are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms. They perform deep analysis of linguistic phenomena.

Statistical Approach: Statistical approaches employ various mathematical techniques and often use large text input to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text input without adding significant linguistic or world knowledge. Machine learning based NLP solutions use this approach.

Connectionist Approach: Connectionist approach develops generalized models from examples of linguistic phenomena. However, connectionist models combine statistical learning with various theories of representation – thus the connectionist representations allow transformation, inference, and manipulation of logic formulae. [1]

Connectionist NLP approach is newer compared to symbolic and statistical approaches. Connectionist NLP work first appeared in the 1960's. For a long time, symbolic approaches dominated the field however, in the 1980's, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real-world contexts. Connectionist approaches also recovered from earlier criticism by demonstrating the utility of neural networks in NLP. [2]

Natural language processing can be used in various application such as translation between languages, dialogues systems (such as Customer Care) and the most important being information extraction (IE). The main goal information extraction is to transform unstructured text into structured (database) representations that can be searched and browsed in flexible ways.

Natural language processing includes following tasks. Some of these tasks can server as real-world application and others might be sub-tasks for other tasks.

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.57

Information Extraction – IE is the process of converting unstructured text into structures or semi-structured form. The unstructured text can be translated to standard databases, which can be queried by users. [3]

Information Retrieval – IR is the process of determining information resources, which are relevant to a query from a collection of resources such as text files, images, videos etc. The most proper form of information-seeking behavior is considered as Information Retrieval. [4]

Relationship Extraction – It is the process of detecting semantic relationships between a set of articles. The co-occurrence of term or their synonyms can be treated as an indicator of relationship between to artifacts. [5]

Co-reference Resolution – Co-reference resolution plays a very important role in information extractions. It is the process of marking up two expressions with the same entity. [6] e.g. Timi is very nice boy, he is very hard working. In this sentence, Timi and hei refers to same person. There are different types of distinctions that can be made out of co-references: Anaphora, Cataphora, Split antecedents, and Co-referring noun phrases.

Named Entity Recognition - Named Entity Recognition is a process of marking up different parts/atoms of a sentence to their respective entities e.g Person Name, Quantity, and Size etc. It has been regarded as an efficient strategy to capture relevant entities for answering different queries. [7]

Part-of-speech tagging – Part-of-speech tagging is the process of tagging each word with its respective part-of-speeches. The part-of-speech tagging is often ambiguous due to different forms of words. The problem of part-of-speech disambiguation can only be solved after solving other problems associated with natural language understanding. [8]

Syntactic Parsing – It is the process of analyzing a sentence by determining the structure of its constituent parts. A parse tree is formed in the process of syntactical parsing. The syntactic parsing plays a very important role in semantic role labeling. [9]

Sentiment Analysis - It is also called as Opinion mining. Sentiment analysis is the process of extracting the subjective information from a piece of text. It is the analysis where the information extracted can be the tone of the author, the inferences that can be made out of that text. [10]

Other tasks may include: Word sense disambiguation, word segmentation, topic segmentation, sentence breaking, morphological segmentation, Discourse analysis, stemming etc.

II. LITERATURE REVIEW

The Natural language processing has variety of applications in real world applications. The Natural language processing is being used extensively in bio-medical field. The National Library of Medicine's MetaMap program is being used extensively for text-mining and creates a standard for indexing of bio-medical terms. [11] There are many other researches which are related to work in bio-medical text mining, such as - cause of death [12], health score [13], smoking status [14][15][16].

NLP is also being applied in fields such as Automated Customer Care service. The other applications of NLP include generating SQL queries from plain text based on synonymous words, [17] the flight schedule query system, [18] improving communications in e-democracy, [19] intrusion detection, [20] text encryption, [21] ontology based natural language processing for in-store shopping, [22] and software requirements specification analysis [23] etc.

In DINLP, Apache UIMA, cTAKES is being used and modified for the needs of domain-independent natural language processing and term extraction system. The background for this paper can be divided into three categories broadly: Apache UIMA, Apache cTAKES and Text analysis. Below is the diagram showing the relationship between Text Analysis, cTAKES and UIMA.

cTAKES (Clinical Text Analysis and Knowledge Extraction Syetem) is a text analysis system developed by Mayo Clinic now being maintained by Apache. It uses Apache's UIMA (Unstructured Information management Architecture) for converting the unstructured text to structured form. The DINLP uses cTAKES for all the text analysis and knowledge extraction.

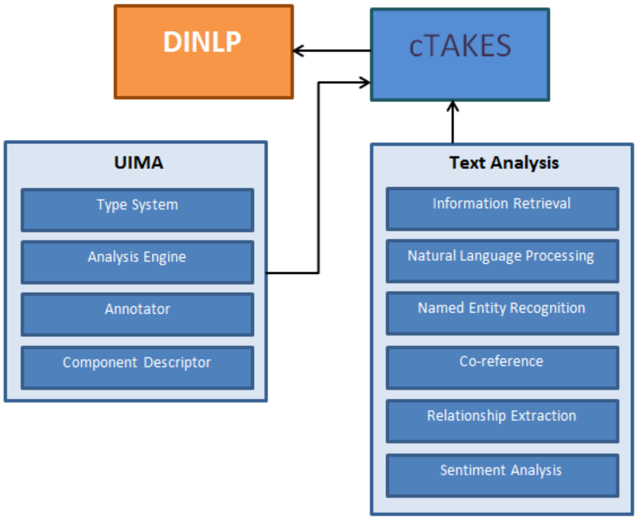

Fig. 1. Relationship between components of DINLP

Text Analysis: The text-based materials are very important source of valuable information and knowledge. There are varying pieces of text which can be treated as Information source such as: Discharge Summaries for Health Care, Accident reports for Road Safety etc. [24] All the sources provide experiment results/summaries as free text which is easily readable by human, but complex for computers to understand. [25] The important components of a general text analysis system are: Information retrieval [26], Natural language processing [27], Named Entity recognition [28], Co-reference [29], Relationship Extraction [30], and Sentiment analysis. The text-analysis is being used in Bio-medical text mining, social media monitoring, National Security and Enterprise business intelligence etc.

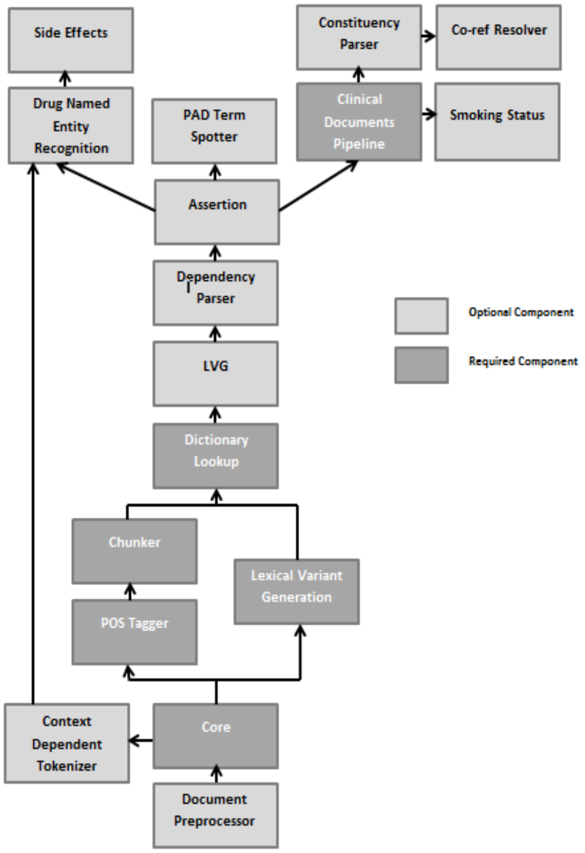

Fig. 2. Components of cTAKES

Apache UIMA: The data in today's world is in unstructured form like in discharge summaries, handwritten notes, research results/reports etc. This unstructured information is very complex because this information is written with no standard structure being followed. The text may contain spelling mistakes, the author's attitude and the main complexity arises when every person has their own terminologies for defining a problem or situation. Due to the increase in such complex unstructured information, we needed an architecture which can convert this unstructured data into structured form in a standardized manner. The new architecture needs to extract the information which can be later related to concepts and events. [31] The UIMA is divided into four main parts: Type System, Analysis Engine, Annotator and CAS. UIMA is also used by: NLM's MetaMap, YTEX and Detect – HAI.

cTAKES: The cTAKES is a modular system of pipelined components combining rule-based and machine learning techniques aiming at information extraction from the clinical narrative. The system is being used to process and extract information from free-text clinical notes for PAD(Peripheral Artery disease). It consists of loosely coupled components and each component has unique capabilities and responsibilities.

III. METHODOLOGY

The system liberally borrows some of the components from the Apache UIMA framework and Apache cTAKES. This section will give a detailed overview of what has been taken from cTAKES and UIMA with information about all components which are modified to suit domain-independent needs.

The Apache UIMA framework is the core of this system. It is built on top of all the components such as their Annotators, type system and Analysis Engines. Below is the overall process flow of this system which is made by modification of cTAKES maintained by Apache foundation. The system is using few of the components provided in cTAKES and also adding some more new components.

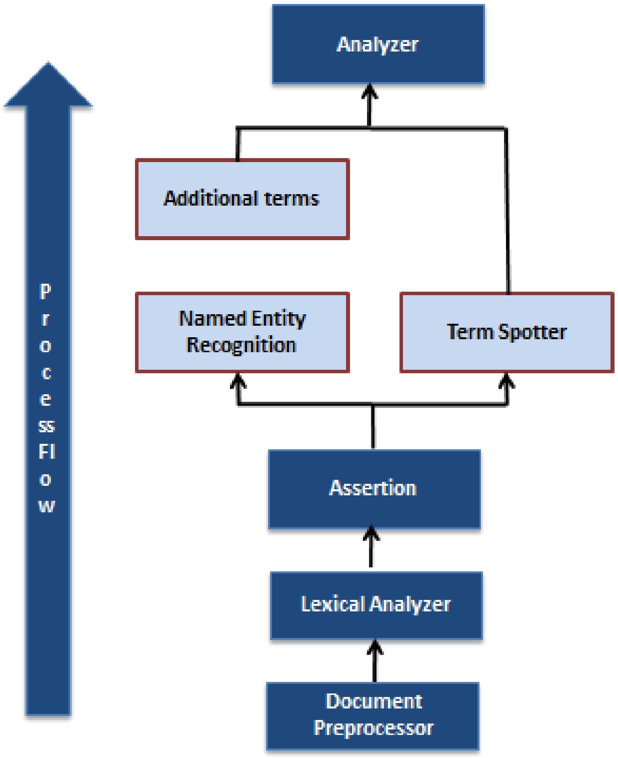

Fig. 3. DINLP process flow

The role and responsibilites and each and every component is as follows:

• Document preprocessor: This component of DINLP was built by integrating separate components of cTAKES. The two components which were borrowed from cTAKES are:

Document Preprocessor and Core. The document preprocessor component helps in converting all the discharge summaries which are in form of Clinical Document Architecture (CDA) for all healthcare related documents. The CDA must conform to DTD. The Core in turn provides two separate components Sectionizer and Tokenizer. The Sectionizer divides a document in different section based on the section markers added in Document Preprocessor step and the Tokenizer converts each section into different tokens for further evaluation. The sentence detector model in cTAKES is derived from a combination of GENIA, PENN Tree Bank and anonymized clinical data.

- Lexical Analyzer: The Lexical Analyzer component of DINLP is built by slight modification and integration of different cTAKES components. This component is built by integrating following component from cTAKES: Lexical Variant Generator (LVG), Context Dependent Tokenizer, Part-of-speech tagger, Chunker, Chunk Adjuster. The LVG generates canonical forms of words for analysis e.g run, ran and running are forms of same lexeme with run as lemma. The CDT creates different annotation with one or more tokens e.g. range annotation, 2-3. The POS tagger assigns each word to its part of speech based on the context in which it is used e.g. store can be used for storing and it can also pertain to a departmental store. The Apache cTAKES POS tagger is built by using three different training data sources: GENIA, Penn Treebank and Mayo part-of-speech corpus. The Chunker creates chunks of noun phrases, verb phrases etc. It is also called as Shallow Parser. The Chunker partitions plain text into sequences of semantically related words. The Chunk Adjuster provides the ability to pseudo-merge different chunks such as Noun Phrase (NP), Verb Phrase (VP), Prepositional Phrase (PP), Adjective Phrase (AP), Adverb Phrase (AdvP).

- Assertion: This component provides a mechanism for examining and documenting the real-world implications for annotations in text. This component asserts whether a given named entity or event is negated, uncertain or conditional. This is taken care by two annotators: Negation Annotator and Context Annotator. For example: In healthcare documents, if "diabetes" is mentioned in some text. The context in which "diabetes" is used is very important. The text might say "The Patient has diabetes" or it may say "The patient's father has diabetes" or it might be used in a generic way "The patient was given a diabetes brochure". So, the Assertion component provides the ability to assert the actual implications of text.

- Term Spotter: Term spotter processes textual extractions which specifically pertains to domain related terms and annoations. For example – in security measures, the terms can be intrusion, plagiarism etc. This component is built by huge modification of cTAKES' PAD Term Spotter which was used to extract textual information for Peripheral Artery Disease. This component extracts the information whether the text in process is pertaining to that domain or not. There are two main dictionaries involved – Primary, which contains all direct domain related terms and Secondary, which contains all the terms reated to secondary information about domain.

- Named Entity Recognition: This component is built on top of the Drug NER component provided by cTAKES. It process text to extract information from it which is provided in form of dictionary. The dictionaries can be either CSV, database or flat file etc.. The component extracts the dictionary information from both narrative text and lists.

The dictionary being used in this component is built with Lucene indexes. It is a very high-performance and scalable text search engine library by Apache Software Foundation. These lucene indexes can search for the drug mentions very fast. The speed is over 150GB/hour on modern hardware. [32]

- Additional terms: This component extracts all the additional data related to the extracted in NER component. Below is the format in which the term and additional information need to be mentioned:

generic term|information 1, information 2,...

e.g. If we need to use this additional component as side-effect extractor for all the drug mentions in text, below will be the format:

Buspirone|Buspar|feeling light-headed, fainting, fast heart rate, uneven heart rate, depressed mood, unusual thoughts, unusual behavior, lack of balance, lack of coordination, drowsiness, dizziness, blurred vision, feeling restless, nausea, upset stomach, sleep problems, insomnia, trouble concentrating.

- Analyzer: This component is very vital to this system, since it plays a very iportant role of putting everything together. This component extracts output from each component and make a result report with all the terms extracted related to COPD diagnoses, medications, lab reports, side-effects etc.

IV. RESULTS

The DINLP was tested for the healthcare domain. Task at hand was of predicting hospital readmission risk for COPD[33][34][35] .The main objective of the test was to extract all the COPD diagnosis and disease related information. The medication list was also provided to test the Named Entity recognition component and additional term component was used to extract side-effect information from all the drug name mentions.

The system is tested by analyzing 1695 discharge summaries' of patient diagnosed with COPD, some of them were a case of readmission related to COPD. The terms, which are extracted, is stored in dictionaries in form of Comma Separated Value files. There are four main dictionaries involved: Primary (COPD diagnoses), Secondary (COPD Anatomical sites), Medications, Side-effects. This section will show detailed analysis reports for each term extracted divided into several sections.

- COPD vs Smoking Status: This graph shows that COPD is very rare in patients who does not smoke. It also shows that for past smokers we found traces of COPD and among current smokers COPD is high percentage. We cannot make any comment about Unknown smoking status

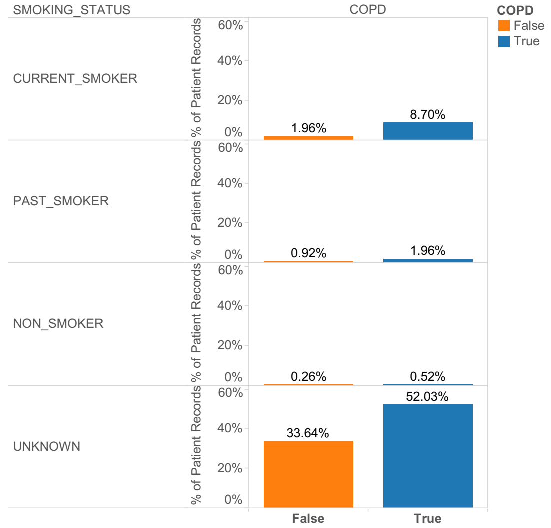

Fig. 4. COPD VS Smoking Status

• COPD VS Medications: This graph shows the relation between COPD and its related medications. This graph gives an idea that almost 8.18% COPD patients take Albuterol while only 2.16% non-COPD patient take Albuterol. The graph also shows that 54.06% of COPD patients are not taking either Albuterol, Ventolin, Xopenex which means that there should be other medicines which are prescribed to COPD patients

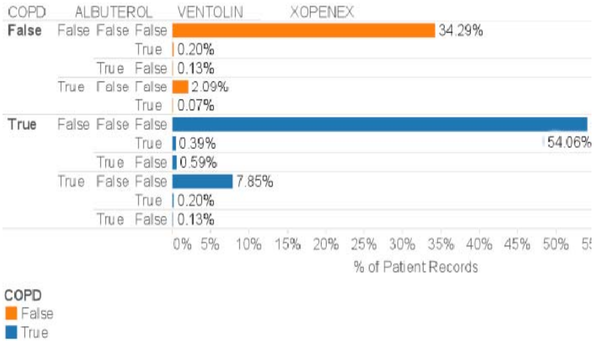

Fig. 5. COPD VS Medications

V. Conclusion

The DINLP is proven to be working fine in the healthcare domain with Chronic Obstructive Pulmonary Disease. In the same way, DINLP can be used in various fields.

References

[1] WebSite http://omarsbrain.wordpress.com/2010/08/12/natural-language-processing-the-big-picture/

[2] Encyclopedia of Library and Information Science, Second Edition -, Volume 3 edited by Miriam Drake

[3] Jim Cowie , Wendy Lehnert, Information extraction, Communications of the ACM, v.39 n.1, p.80-91, Jan. 1996

[4] Belkin, N.J. (1993) Interaction with texts: Information retrieval as information-seeking behavior. In: Information retrieval '93. Von der Modellierung zur Anwendung. Konstanz: Universitaetsverlag Konstanz, 55-66.

[5] Qiang Song , Yousuke Watanabe , Haruo Yokota, Relationship extraction methods based on co-occurrence in web pages and files, Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, December 05-07, 2011, Ho Chi Minh City, Vietnam

[6] Soon W, Lim D, Ng H (2001) A machine learning approach to co-reference resolution of noun phrases. J Comput Linguist 27(4):521–544

[7] J. J. Jung. Online named entity recognition method for microtexts in social networking services: A case study of twitter. Expert Systems with Applications, Volume 39, Issue 9, July 2012, Pages 8066-8070

[8] S. Abney. Part-of-Speech Tagging and Partial Parsing. Corpus-Based Methods in Language and Speech Processing Text, Speech and Language Technology Volume 2, 1997, pp 118-136.

[9] Vasin Punyakanok , Dan Roth , Wen-tau Yih, The importance of syntactic parsing and inference in semantic role labeling, Computational Linguistics, v.34 n.2, p.257-287, June 2008

[10] Yoonjung Choi , Youngho Kim , Sung-Hyon Myaeng, Domain-specific sentiment analysis using contextual feature generation, Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, November 06-06, 2009, Hong Kong, China

[11] Alan R Aronson AR, Lang FM., "An overview of MetaMap: historical perspective and recent advances," J Am Med Inform Assoc, 17:229-236, 2010

[12] Shah AD, Martinez C, Hemingway H., "The Freetext Matching Algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records," BMC Medical Informatics and Decision Making, 12:88, 2012

[13] Pakhomov S, Shah N, Hanson P, Balasubramaniam S, Smith SA,

[14] "Automatic quality of life prediction using electronic medical records," AMIA Annu Symp Proc, 2008:545–549, 2008.

[15] Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG, "Mayo clinic NLP system for patient smoking status identification," J Am Med Inform Assoc, 15:25-28, 2008.

[16] Uzuner O, Goldstein I, Luo Y, Kohane I, "Identifying Patient Smoking Status from Medical Discharge Records," J Am Med Inform Assoc, 15:14–24, 2008

[17] Siasar djahantighi, F, Norouzifard, M. ; Davarpanah, S.H. ; Shenassa, M.H. Using natural language processing in order to create SQL queries. Computer and Communication Engineering, 2008. ICCCE. 600 – 604. 2008.

[18] Hamza, M.A.B.M. ; Ahmad, A.M. Flight schedule query system based on natural language processing. Research and Development, 2002. SCOReD. 80 – 82. 2002

[19] Carenini, M., Whyte, A., Bertorello, L., Vanocchi, M. Improving Communication in E-democracy Using Natural Language Processing. Intelligent Systems, IEEE (Volume:22 , Issue: 1). 20 – 27. 2007.

[20] Stone, A. Natural-Language Processing for Intrusion Detection. Computer (Volume:40 , Issue: 12). 103 – 105. 2007.

[21] Xianghe Jing, Yu Hao, Huaping Fei, Zhijun Li. Text Encryption Algorithm Based on Natural Language Processing. Multimedia Information Networking and Security (MINES). 670 – 672. 2012.

[22] Janzen, S., Maass, W. Ontology-Based Natural Language Processing for In-store Shopping Situations. Semantic Computing, 2009. ICSC '09. 361 – 366. 2009.

[23] Fatwanto, A. Software requirements specification analysis using natural language processing technique. QiR (Quality in Research). 105 – 110. 2013.

[24] Clarizia, F., De Santo, M. ; Napoletano, P. A Probabilistic Method for Text Analysis. Intelligent Systems Design and Applications, 2009. ISDA '09. 10.1109/ISDA.2009.26.

[25] M. Krallinger, A. Valencia. Text-mining and information-retrieval services for molecular biology. Genome Biology 2005, 6:224. doi:10.1186/gb-2005-6-7-224

[26] Sneiderman C, Demner-Fushman D, Fiszman M, Ide N, Rindflesch T: Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc 2007, 14:772-780

[27] Ah-Hwee Tan. Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases KDAD'99, page 65-70. (1999)

[28] J. J. Jung. Online named entity recognition method for microtexts in social networking services: A case study of twitter. Expert Systems with Applications, Volume 39, Issue 9, July 2012, Pages 8066-8070

[29] Soon W, Lim D, Ng H (2001) A machine learning approach to co-reference resolution of noun phrases. J Comput Linguist 27(4):521–544

[30] Ce Gao,Yixu Song, Peifa Jia. A Fast Events Relationship Extraction Method Based on Semi-CRFs. Knowledge Acquisition and Modeling, 2009.

[31] Michael Thomas Egner, Markus Lorch and Edd Biddle. UIMA GRID: Distributed Large-scale Text Analysis, 2007, IEEE

[32] Apache Lucene website. http://lucene.apache.org/core/

[33] Ravi Behara, Ankur Agarwal, Faiz Fatteh, Borko Furht, Predicting Hospital Readmission Risk for COPD Using EHR Information, "Handbook of Medical and Healthcare Technologies", Springer, 2014

[34] Borko Furht, Ankur Agarwal (Editors), Handbook of Medical and Healthcare Technologies, Springer Publication, October 2013, ISBN: 978-1-4614-8494-3

[35] Ravi Behra, Faiz Fatteh, Piyush Jain, Prvatha Laxmi, Ankur Agarwal, "A Prognostic COPD Healthcare Management System", IEEE International Conference on Systems Engineering, 2013, Orlando, Florida