Text Mining used in the Field of Cancer Detection

Meena Preethi. B
Assistant Professor
Department of BCA & MSC.SS
Sri Krishna Arts and Science College, Coimbatore-641008

Sruthi. R
Student
Department of BCA & MSC.SS
Sri Krishna Arts and Science College, Coimbatore-641008

Darshna. R
Student
Department of BCA & MSC.SS
Sri Krishna Arts and Science College, Coimbatore-641008

Abstract

Cancer is a malignant disease that has caused millions of human deaths. Text mining can help researchers discover hidden rules and relationships between documents so advanced cancer research. In this paper, we analyze the properties of text mining and cancer research documents. We discussed the research directions of text mining in cancer research with examples of systems and tools. In conclusion part, we talked about future way the text mining development.

Keywords: Text Mining, Cancer Domain, Risk Assessment

I. INTRODUCTION

Cancer is a malignant disease that has caused millions of human deaths. In 2012, about 14.1 million new cancers occurred globally, and caused about 8.2 million deaths [1], which is equivalent 14.6% of all human death [2]. Biomedical researchers spent a lot time and effort and time trying to find the way to cure cancer. Also, pharmaceutical and biopharmaceuticals companies invested heavily on the oncology studies. Fig. 1 presents the general framework in which TM is used in the clinical setting. Different types of cancer (e.g. breast cancer) will define a specific domain in which TM is used. It will determine the choice of available text data (e.g. mammography reports). Data interpretation, either by human experts or computers, naturally requires relevant knowledge in given domain. In recent years, biomedical text mining has increased in popularity. Techniques have been developed to assist, for example, the extraction of documents, databases, dictionaries, ontologisms,

Summaries and specific information (e.g. interactions between proteins and genes, novel research hypotheses) from relevant literature [4]. In this paper we present a new, fully integrated text mining system designed to support the complex and highly literature dependent task of chemical health risk assessment. This task is critical because chemicals play an important role in everyday life and their potential risk to human health must be evaluated. With thousands of chemicals introduced every year, many countries worldwide have established increasingly strict laws governing their production and use. For example, the recent European Union Registration, Evaluation, Authorization and Restriction (REACH) legislation [5] requires that all chemicals manufactured or imported in large quantity must undergo thorough risk assessment. The assessment of large numbers of chemicals is easier said than done. Using the currently available methodology, it takes up to two years to assess a single chemical [6]. Although the development of a completely novel system for toxicity testing may help to improve the efficiency of chemical assessment in the long term [7], there is a pressing need to improve the state of the art in the short to medium term.

Chemical risk assessment is a complex process consisting of several component stages. The chief foremost component is typically an extensive review and analysis of the available scientific data on the chemical in question. This review focuses on any data of potential relevance – not only human data, but also animal, cellular (in vitro) and other mechanistic data [8]. The primary source for this data is scientific peer reviewed literature.

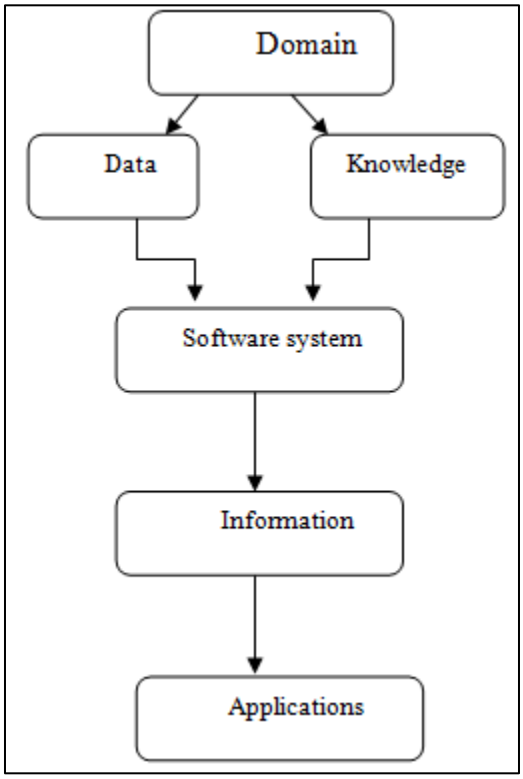

Fig. 1: Text mining framework for clinical applications

II. CANCER RISK ASSESSMENT

Cancer risk assessment is the process to identify the relationship between chemical and exposure from existing published evidence [17]. In worldwide, more and more evidence shows the link between environmental chemicals and cancer and government legislations becomes tight.

The critical tool used by authorities in making decisions on exposure limits is Risk Assessment (RA). Cancer RA involves examining existing published evidence to determine the relationship between exposure to a substance and the likelihood of developing cancer from that exposure. From the perspective of TM, cancer RA provides a suitably complex use case for tackling the timeliest problems in the field.

III. CANCER DOMAINS

Cancer is an umbrella term for diseases characterized by excessive cellular division and proliferation. Cancer is no longer viewed as a single disease or even a single collection of diseases. Cancer may start developing in any of over 60body organs and is usually named after the affected organ (e.g. breast cancer). Each organ consists of different cell types that may be affected by cancer, so several cancer types may affect each organ (e.g. ductal carcinoma and lobular carcinoma are different types of breast cancer). There are more than 200 types of cancer having different causes, symptoms and treatments.

The literature review revealed that most of the articles on TM for cancer have been annotated with a MeSH term that identifies neoplasm by anatomical site, where neoplasm refers to autonomous tissue growth in which the malignancy status has not been established and for which the transformed cell type has not been specifically identified. Both NCI The-saurus [9] and MeSH [10] organize neoplasms by anatomical site.

IV. TEXT DATA PROCESSING

In this section we focus on specific text processing tasks together with a review of methods and techniques used to solve them in the cancer domain. We differentiate between four major NLP tasks: named entity recognition (NER), information extraction (IE), text classification and information retrieval (IR). In order to assess the feasibility and compare different approaches, we first describe how the systems sup-porting these NLP tasks can be evaluated.

A. Named Entity Recognition:

NER identifies and classifies words and phrases into pre-defined categories such as diseases, symptoms and drugs. NERis used mostly as a vehicle for feature extraction in order to support more complex NLP tasks such as IE, text classification and IR. Most approaches reviewed in this article relied on dictionary-based NER methods to recognize cancer types and gene names. The

overwhelming majority used MetaMapto recognize concepts from UMLS [11], usually focusing on specific classes (or semantic types as they are called in the UMLS documentation) of concepts.

The biomedical domain exhibits high degree of terminological variation, which stems from the ability of a natural language to name a single entity in different ways. It has been estimated that approximately one third of term occurrences are variants [12]. The cancer domain is no exception, where various synonyms for each cancer type exist.

Kang et al. [13] used MetaMap as a baseline and demon-strated how its performance on disease names (includingcancers) can be further improved using a rule-based approach. Kang et al. combined shallow parsing with a number of rules that adjust noun phrases and feed them back into the nor-malisation process to check whether they refer to known entities. Their error analysis highlighted problems associated with cancer type recognition often due to coordination. Their approach included coordination resolution in which part-of-speech and chunking information was used to reformat the coordination phrase such as colorectal, endometrial and ovar-ian cancers and recognize ovarian cancers, colorectal cancers and endometrial cancers as separate entities.

Fang et al. developed a cancer name entity recognizer apart of their MeInfo Text system for mining gene methylation and cancer association information [14]. They combineda cancer dictionary and regular expression patterns. The dictionary of cancer names including their abbreviations was compiled from the previous version of the system.

Table – 1

System	*Brief introduction*
ABNER	*ABNER is a software tool for molecular biology text analysis. It uses linear-chain conditional random fields approach with orthographic and contextual features*
GENIATagger	*The GENIA tagger is specifically tuned for biomedical text such as MEDLINE abstracts. It is a useful pre-processing tool for information extraction from biomedical documents*
LingPipe	*LingPipe provides three generic, trainable chunkers to carry on named entity recognition. LingPipe can be used to identify biomedical entities such as genes, organisms, malignancies, and chemicals*
Yapex	*Yapex is a rule-based system named entity recognition system that utilizes lexical and syntactic analysis to identify protein names*

B. Information Extraction:

IE selects specific facts about pre-specified types of entities and relationships of interest. For example, the 2009 i2b2medication extraction challenge focused on the extraction of medication-related information including: medication name(m), dosage (do), mode (mo), frequency (f), duration (du) and reason (r) from hospital discharge summaries. In other words, free-text medical records needed to be converted into a structured form by filling a template (a data structure with the predefined slots) with the relevant information extracted (slotfillers). In this task, the sentence "In the past two months, she had been taking Ativan of 3–4 mg q.d. for anxiety." should be converted automatically into a structured form as follows [15]:
m="ativan" || do="3–4 mg" || mo="nm" || f="q.d." || du="two months" || r="for anxiety" where nm indicates that particular information was not mentioned.

C. Text Classification:

IE converts free text data into structured information, which adds significant value to the data in terms of automated analyses that can then be performed over semantically typed and structured data. In layman terms, one could view IE as a conversion of a Word document into an Excel spreadsheet, which makes complex statistical calculations only a click away. However, IE only applies to explicitly stated information. More value can be gained by inferring additional information that is not explicitly articulated in the original text. This can be achieved using text classification, which uses features extracted from text (e.g. using NER or more advanced IE) to map text (e.g. sentence, paragraph or most often a whole doc-ument) into one or more classes from a predefined scheme.

V. TEXT MINING SYSTEMS

A table in the Supplementary material provides a summary of the systems described in this section focusing on particular NLP tasks: named entity recognition, information extraction, text classification and information retrieval. Here we provide a more detailed overview of two more generic NLP systems thathave been developed for and/or tested in the cancer domain.

A. MedLEE:

The goal of the MedLEE (Medical Language Extraction and Encoding) system, developed at Columbia University in collaboration with the City University of New York, is to extract, structure and encode clinical information in free text patient reports so that it can be further exploited by subsequent auto-mated processes. Originally, MedLEE was designed to process radiological reports of the chest to detect patients suspicious for tuberculosis [16].

MedLEE was used to code 889,921reports on 251,186 patients. Using a set of 150 manually coded reports as a gold standard, sensitivity of 81% and specificity of 99% were reported. A total of 24 clinical conditions (dis-eases, abnormalities and clinical states) were the subject of this study. We believe that focusing on lung cancer alone would allow for finer tuning of the underlying lexical and domain-specific knowledge, which would be reflected in better sensitivity.

B. cTAKES:

(clinical Text Analysis and Knowledge Extraction System), a generic NLP system developed at the Mayo Clinic, is tailored to the clinical domain and can add rich linguistic and semantic annotations to the narrative found in EMRs [18].It has been designed to be scalable, comprehensive, modular, extensible and robust to meet the rigours of clinical research. cTAKES consists of the following NLP modules: sentence boundary detector, tokenizer, normalizer, POS tagger, shallow parser and NER annotator.

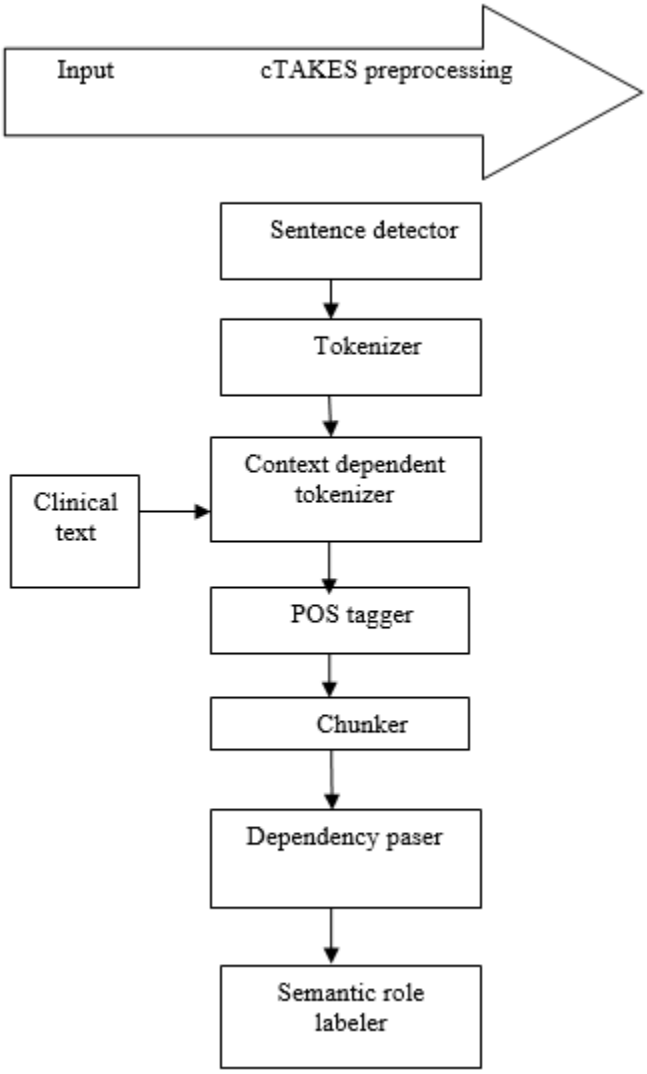

Fig. 2: cTAKES processing

VI. CONCLUSION

Due to its powerful information retrieval, knowledge discovery ability from unstructured text, Text mining has already played a big role for cancer research. Many researches and systems have been developed to help biomedical and biomedicine researchers. The genes or proteins studied by various research groups by carefully analyze their published research articles to identify the molecules they reported as biological biomarkers of breast cancer. Interestingly, we realized that researchers have reported interest in a variety of genes over time and even based on the country where the research is conducted. The tool can also be stretched to maintenance analysis of the scientific data and the subsequent writing of risk assessment reports.

REFERENCES

[1] World Health Organization, "World Cancer Report 2014," 2014.
[2] WHO, "The top 10 causes of death," May 2014 [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310/en/.[Accessed 12 May 2015].
[3] D. Rebholz-Schuhmann, H. Kirsch, and F. Couto, "Facts from text— Is text mining ready to deliver?" Plos Biology, vol. 3, no. 2, p.65, 2009.
[4] Ananiadou S, McNaught J (2006) Text Mining for Biology And Biomedicine. Norwood, MA: Artech House, Inc.
[5] http://ec.europa.eu/environment/chemicals/reach/legislation_en.htm. Accessed 2012 February 17.
[6] Judson R, Richard A, Dix DJ, Houck K, Martin M, et al. (2009) The toxicity data landscape for environmental chemicals. Environmental Health Perspectives 117: 685–695.

[7] US National Academy of Science (2007) Toxicity testing in the 21st Century: A vision and a strategy. http://dels-old.nas.edu/dels/rpt_briefs/Toxicity_ Testing_final.pdf Accessed 2012 February 17.
[8] US Environmental Protection Agency (EPA) (2005) Guidelines for Carcinogen Risk Assessment. http://www.epa.gov/cancerguidelines/. Accessed 2012 February 17.
[9] US National Library of Medicine, Medical Subject Headings(MeSH), 2013 http://www.nlm.nih.gov/mesh/
[10] US National Library of Medicine, UMLS Terminology Services, 2013 https://uts.nlm.nih.gov/
[11] A.R. Aronson, Effective mapping of biomedical text to theUMLS Metathesaurus: the MetaMap program, in: AmericanMedical Informatics Association, 2001, pp. 17–21.
[12] C. Jacquemin, Spotting and Discovering Terms ThroughNatural Language Processing, MIT Press, Cambridge, MA,2001
[13] N. Kang, B. Singh, Z. Afzal, Mulligen EMv, J.A. Kors, Usingrule-based natural language processing to improve diseasenormalization in biomedical text, J. Am. Med. Inform. Assoc.20 (2013) 876–881.
[14] Y.-C.C. Fang, P.-T.T. Lai, H.-J.J. Dai, W.-L.L. Hsu, MeInfoText2.0: gene methylation and cancer relation extraction from biomedical literature, BMC Bioinformatics 12 (2011)47.
[15] I. Spasi´c, F. Sarafraz, J.A. Keane, G. Nenadi´c, Medicationinformation extraction with linguistic pattern matching andsemantic rules, J. Am. Med. Inform. Assoc. 17 (2010) 532–535.
[16] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, Ageneral natural-language text processor for clinicalradiology, J. Am. Med. Inform. Assoc. 1 (1994) 161–174.
[17] U. E. P. A., "Guidelines for carcinogen risk assessment 2005," [Online] Available: http://www.epa.gov/iris/cancer032505.pdf
[18] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C.Kipper-Schuler, et al., Mayo clinical Text Analysis andKnowledge Extraction System (cTAKES): architecture,component evaluation and applications, J. Am. Med. Inform.Assoc. 17 (2010) 507–513.