NOVEL APPLICATIONS OF NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TO EXTRACT INFORMATION FROM CLINICAL TEXT AND AUTOMATE CANCER STAGE COLLECTION IN A CENTRAL CANCER REGISTRY

by

Abdulrahman Khalifa AAlAbdulsalam

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2018

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Abdulrahman Khalifa AAlAbdulsalam**

has been approved by the following supervisory committee members:

Wendy W. Chapman , Chair(s) **12/14/2017**
Date Approved

Stéphane M. Meystre , Member
Date Approved

Jennifer H. Garvin , Member
Date Approved

Guilherme Del Fiol , Member **12/14/2017**
Date Approved

Stephen S. Piccolo , Member **12/14/2017**
Date Approved

by **Wendy W. Chapman** , Chair/Dean of

the Department/College/School of **Biomedical Informatics**

and by **David B. Kieda** , Dean of The Graduate School.

ABSTRACT

The primary objective of cancer registries is to capture clinical care data of cancer populations and aid in prevention, allow early detection, determine prognosis, and assess quality of various treatments and interventions. Furthermore, the role of cancer registries is paramount in supporting cancer epidemiological studies and medical research. Existing cancer registries depend mostly on humans, known as Cancer Tumor Registrars (CTRs), to conduct manual abstraction of the electronic health records to find reportable cancer cases and extract other data elements required for regulatory reporting. This is often a time-consuming and laborious task prone to human error affecting quality, completeness and timeliness of cancer registries.

Central state cancer registries take responsibility for consolidating data received from multiple sources for each cancer case and to assign the most accurate information. The Utah Cancer Registry (UCR) at the University of Utah, for instance, leads and oversees more than 70 cancer treatment facilities in the state of Utah to collect data for each diagnosed cancer case and consolidate multiple sources of information.

Although software tools helping with the manual abstraction process exist, they mainly focus on cancer case findings based on pathology reports and do not support automatic extraction of other data elements such as TNM cancer stage information, an important prognostic factor required before initiating clinical treatment.

In this study, I present novel applications of natural language processing (NLP) and machine learning (ML) to automatically extract clinical and pathological TNM stage information from unconsolidated clinical records of cancer patients available at the central Utah Cancer Registry. To further support CTRs in their manual efforts, I demonstrate a new approach based on machine learning to consolidate TNM stages from multiple records at the patient level.

To my beloved parents, my wife and children . . .

CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

In the name of Allah (God), the Merciful, the Beneficent. First and foremost, I would like to express my deep gratitude to my wife, Adhari, who has supported me and spent countless hours listening to my research. She did all of this despite her very busy schedule working on her graduate school and taking care of our three children: Zainab, Yusuf and Maryam. I am very grateful for her and my parents for their warm feelings and persistent encouragement. I am very proud of them and words are not enough to describe their kindness and love.

This work would not have been possible without the generous time of my mentor, Stéphane Meystre. He spared no effort in helping me throughout my entire PhD study even after he left the department for his new position in South Carolina. I would like to specially thank Wendy Chapman for agreeing to become my new committee chair. I feel very lucky to get two great mentors during my doctorate study. I am very thankful for the rest of my committee members: Jennifer Garvin, Guilherme Del Fiol, Mike Conway and Stephen Piccolo for their critical assistance and thoughtful feedback which helped shape the final research in this manuscript.

I am very thankful for the staff at the Utah Cancer Registry for their collaboration and considerable assistance in completing major research work described in this dissertation. Special thanks to their director Carol Sweeny and the rest of the team: Marjorie Carter, Kim Herget, SuAnn McFadden, Jacque Clarken, Kay McCandless, Carrie Bateman, and Valerie Otto. Special thanks to Andrew Redd for providing necessary statistical analysis for the results reported in this dissertation.

Finally, I would like to thank all of faculty, the administrative staff and students at the Department of Biomedical Informatics for being a wonderful and fun community to work with, especially, Barbara Saffel, JoAnn Thompson, Tahanie Branz, Diane Walker, Valliammai Chidambaram and Aly Khalifa. And, last but not least, very special thanks to my friend, Cody Hill, for proof reading early drafts of this dissertation.

CHAPTER 1

INTRODUCTION

The recent adoption of Electronic Health Record (EHR) systems across many healthcare institutions significantly improved the ability to capture, store and organize clinical data, enabled many clinical applications, and facilitated data reuse for medical research.[1,2] However, the amount of clinical data is increasing dramatically and is mostly realized in unstructured narrative text format that is difficult to process automatically by computers and derive useful knowledge.[3] In addition, being the preferred medium of documentation and communication by clinicians, narrative clinical text is predominantly the most abundant form and contains the greatest amount of information in the EHR.[4–6]

Natural Language Processing (NLP) is an emerging area of research that focuses on processing and analyzing free text written in human language using rule-based algorithms and state of the art in statistical Machine Learning (ML). NLP has enjoyed a great amount of progress within the past decade, leading to many successful applications within the medical domain.[7,8] Earlier attempts included efforts aimed at automatically acquiring diagnoses from radiology and imaging reports[9,10], collecting problem lists of patients.[11,12], extracting medications and clinical findings from clinical narratives[13,14], and finding personal data such as names, addresses and phone numbers for the purpose of deidentifying clinical records to protect patient privacy and allow sharing and exchange of clinical information.[15]

Success in earlier attempts has motivated further research into far more complex problems which heavily depend on information extracted from clinical text. Recently, for instance, researchers investigated the utility of NLP and ML to to classify patients as to whether they have a certain condition based on their past medical records[16–18] and to find patients satisfying specific clinical criteria for medical research and clinical trial recruitment.[19] More ambitious studies involve automatically summarizing longitudinal patient records either visually or textually which require analysis at a much deeper level based on temporal information and semantic relations between concepts.[20, 21]

NLP applications based on statistical machine learning require access to human experts to create annotated corpora that can be used as a gold standard for the development and evaluation of algorithm in a supervised learning framework. Obtaining manual annotations from experts is difficult and expensive especially in the medical domain where strict privacy laws are in place and clinicians' time is scarce and costly.[22]

Throughout the past decade researchers have developed and shared numerous NLP tools and resources that can be readily used to process free text especially for the English language. While the majority of contributions focus on text written for the general domain such as news articles, there is a growing body of researchers who have tackled clinical text and associated medical applications.[23] The availability of many existing NLP resources and associated corpora in the medical domain was largely possible because of the recent uptake in NLP shared tasks which is instrumental in advancing the state-of-the-art in NLP technology.[22] In this study, we leverage existing resources and corpora developed in NLP challenges or shared tasks and apply them to novel NLP applications.

1.1 Background and Significance

Cancer is the second leading cause of death in the United States and recently became the leading cause of death in 21 states, surpassing heart diseases. About 600,920 cancer deaths were estimated to occur in 2017, which is about 1,650 people per day.[24] The overall 5-year survival rate for cancer is estimated at 69% notwithstanding the fact that this is an overestimated figure because of an increased rate of screening (over-diagnosis).[25] The total expenditure for cancer care in the United States was estimated to have reached $125 billion in 2010.[26] The burden of cancer on public health has mobilized national and international institutions to develop strategies to combat, prevent and control cancer.[27, 28]

Cancer registries are a vital resource in the fight against cancer, paving the way for access to critical clinical care information at the population level. Various cancer clinical care data elements are captured by local and national registries. The role of cancer registries is paramount in supporting cancer epidemiological studies and medical research, in particular estimating cancer incidence and survival rates at the population level. The data collected serve many objectives including ultimately the control of cancer and improvement of patient care.[28] In the United states, both the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program and Center for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR) are responsible for the support and management of cancer registries at the national level. Utah Cancer Registry (UCR) is a member of SEER program which collects primarily incidence and survival data .[29] The North American Association of Central Cancer Registries (NAACCR) maintains data dictionaries and defines coding standards used by cancer registries in the United States and Canada.[30]

Certified Tumor Registrars (CTR) are tasked with identifying reportable cancer cases and manually performing coding of data required for cancer registries. Existing cancer registry departments depend on CTRs to manually curate the electronic records, find reportable cancer cases, and extract all associated data elements such as date of diagnosis, staging information and treatment. This is often a time-consuming and laborious process that is prone to human error and affects quality, completeness and timeliness of cancer registry data.[31] Coding guidelines and data standards are essential to standardization across different registries. Given the overwhelming amount of information contained in these manuals and the complexity of the coding rules for each cancer site, registrars are required to undergo 3 years of training by completing an associate degree and undertaking a clinical practicum before they can become certified for the task.[32] Although computer software helping registrars with their tasks exist, it is mostly limited to cancer case findings and depends on availability of pathology reports and, to the best of our knowledge, do not extend their operations to aid in the extraction of other vital data elements such as cancer stages.[33]

Cancer registry databases suffer from incomplete and slow data reporting due to the manual and laborious process used for data collection. A study based on surveys conducted across European cancer registries as part of the EUROCOURSE project and covering a population of more than 280 million people found that the median time to complete case ascertainment for the relevant year was 18 months, with an additional 3-4 months to publish data to national databases.[34,35] Though delay in completion is primarily related to clinical processes to evaluate the patients extent of disease, reduced time to ascertainment is very desirable. The Utah cancer registry has a similar time-lag with consolidation of cases diagnosed in 2014 only completed in 2016.

Cancer stage information is critical for assessing prognosis and selection of treatment plans and guidelines require staging before initiating any treatment.[36] The American Joint Committee on Cancer (AJCC) manual specifies criteria (known as TNM) for staging each cancer site depending on tumor characteristics (T), number and location of lymph nodes involvement (N), and metastatic nature (M). The AJCC TNM manual is a dynamic resource that is evolving and continually revised to incorporate the most recent cancer knowledge. For instance, for the upcoming 8th edition of the AJCC manual, new factors will be included to determine T stage such as biological markers that are associated with specific cancer sites. Therefore, in addition to laborious manual effort required to abstract cancer stage, there is a learning curve for registrars when assigning stage information due to the dynamic nature of the coding instructions.

Natural Language Processing coupled with Machine Learning are promising technologies to increase the efficiency of cancer registry data abstraction processes. In the domain of cancer, several studies showed effectiveness of NLP and ML to mine the electronic health record for cancer-related information from a variety of report types[37] and automatically discover reportable cancer cases based on analysis of pathology records.[33] We hypothesize that NLP and ML will be effective technologies in aiding human registrars to assign cancer stage based on analysis of multiple unconsolidated records received from reporting hospitals at central cancer registries.

1.2 Research Aims

The overall objective of this research study is to develop novel applications of Natural Language Processing (NLP) and Machine Learning (ML) to automatically extract cancer-related information from records collected at central cancer registry and determining the AJCC TNM cancer staging for each patient. Most previous studies have focused on extracting the AJCC TNM stage information exclusively from pathology reports. This would not support cancer registry efforts adequately since clinical staging is assigned prior to initiation of treatment and many patients do not immediately undergo resection of their tumor and pathology examination. Furthermore, the collection of the AJCC TNM stages from different registry sources requires the additional laborious and time-consuming task of consolidating multiple data at the central cancer registry. In this study, we demonstrate the best classification approach using machine learning to automate cancer stage consolidation from multiple records.

1.2.1 AIM I

1.2.1.1 Justification

NLP applications are usually built from multiple processing components aligned in a pipeline fashion where each component analyzes its input and delivers output to the next component. Existing NLP components and tools can be used to build prepossessing pipeline to transform narrative text to structured units and assign features for further downstream analysis. Adapting and Reusing existing NLP components could substantially reduce development efforts while maintaining a baseline with good performance.

1.2.1.2 Specific Aim I.1

Measure accuracy of reusing existing NLP components and terminology resources to extract risk factors from textual discharge summaries of diabetic patients.

1.2.1.3 Specific Aim I.2

Demonstrate that a baseline system constructed from existing NLP components developed for similar previous tasks can achieve good performance without extensive feature engineering or retraining of modules.

1.2.1.4 Significance

This aim is completed as part of a study[38] for the 2014 i2b2 cardiovascular risk factor identification task[39] using discharge summaries of diabetic patients. The study was an initial preliminary work to experiment with existing clinical NLP resources such as cTAKES[40] and the potential to be adapted as baseline for a new task in finding various clinical information from clinical notes. Much of the NLP pipeline can be reused for other tasks especially the prepossessing components such as sentence detection, tokenization, finding sections, lemmatization (finding word root) and UMLS concept mapping. The study showed that relying on existing NLP resources and adapting them for new tasks helped speed up the development efforts considerably (1-2 months) while maintaining good performance. The proposed NLP system achieved results comparable to the best systems submitted for the task and scored an overall F1-measure of 87.47% in the task which included a total of 49 submissions from 20 teams.

1.2.2 AIM II

1.2.2.1 Justification

Recent NLP applications adopt statistical machine learning frameworks to solve information extraction tasks such as named entity recognition, and relations finding. There are many statistical machine learning algorithms that have been proposed in the literature. Among the most prominent are Support Vector Machine (SVM) and Conditional Random Fields (CRF) algorithms. The SVM approach transforms the input into multidimensional vector space representation and finds the separation line or plane with maximum marginal distance between learning instances. The CRF approach uses Markov chain processes to find the sequence of class labels with highest probability given the observed sequence of input. Both algorithms depend heavily on carefully selected feature representations to improve performance. The best machine learning approach and features for a given NLP task are not obvious upfront and closely examining each approach for a specific task can be useful.

1.2.2.2 Specific Aim II.1

Evaluate accuracy of SVM and CRF structured machine learning approaches to automatically extract cancer-related information, time expressions and relations between them from clinical records of Colorectal Cancer patients.

1.2.2.3 Specific Aim II.2

Compare performance of SVM and CRF and provide future direction for further improvement for similar tasks.

1.2.2.4 Significance

This aim was completed as part of a study[41] for the 2016 Clinical TempEval task.[42] The study compared two popular machine learning approaches in information extraction: sequential classification based on Conditional Random Field (CRF) and large margin classification based on Support Vector Machine (SVM). The results showed that the CRF-based approach slightly outperformed the SVM-based system and that an ensemble-based strategy where predictions from multiple classifiers are combined could yield better results for the time expressions extraction subtask. Our submissions achieved competitive results in each subtask with an F1 score of 75.4% for TIMEX3 subtask, F1 score of about 89.2% for EVENT subtask, F1 score of 84.4% for event relations with document time (DocTimeRel), and F1 score of 51.1% for narrative container (CONTAINS) relations subtask.

1.2.3 AIM III

1.2.3.1 Justification

Central cancer registries receive multiple records from different sources for each newly diagnosed cancer case. Cancer tumor registrars manually perform chart abstraction to find relevant information to stage each cancer case and carry out coding of other data elements required for regulatory reporting. Due to increasing load on the registrars and personnel costs, there is a pressing need to automate the manual abstraction process and support registry's work to improve timeliness and quality of data. NLP and machine learning technologies could be used to build tools for automatic extraction and classification of relevant cancer stage information from clinical text available at records in a central cancer registry.

1.2.3.2 Specific Aim III.1

Develop a reference standard composed of human annotated TNM stage mentions found in text fields of sample records from the Utah Cancer Registry for colon, lung and prostate cancer cases. The reference standard was used for training and evaluation of a new NLP system.

1.2.3.3 Specific Aim III.2

Evaluate accuracy of NLP to automatically extract TNM stage mentions from text in the reference standard and compare against the manual annotations performed by human registrars.

1.2.3.4 Specific Aim III.3

Evaluate accuracy of a CRF-based machine learning approach to classify TNM mentions to clinical or pathological stages.

1.2.3.5 Significance

Automated extraction and classification of TNM stage mentions from unstructured text fields within records at Utah Cancer Registry achieved high accuracy when compared to manual human annotations from the reference standard. The automated extraction using NLP achieved very high sensitivity of about 95.5%–98.4% across the three cancer sites while automatic classification of TNM mentions using the CRF approach achieved sensitivity of about 83.5%–87%.

1.2.4 AIM IV

1.2.4.1 Justification

Consolidation of multiple cancer stages for each cancer case is performed manually by registrars at each central cancer registry. This process has become time-consuming and more expensive with rising cancer cases diagnosed each year. Automated consolidation using machine learning could potentially support registrars' manual effort and reduce costs while maintaining good data quality.

1.2.4.2 Specific Aim IV.1

Develop and evaluate accuracy of machine learning algorithms for consolidating multiple cancer TNM stages.

1.2.4.3 Specific Aim IV.2

Validate the performance of machine learning algorithms for TNM stage consolidation through comparison to the consolidation decisions made by cancer registrars for three cancer sites: colon, lung and prostate.

1.2.4.4 Specific Aim IV.3

Evaluate accuracy of deriving a cancer stage group for each case using TNM stages consolidated by the machine learning algorithm.

1.2.4.5 Significance

Automatic consolidation of cancer stages using machine learning for the cancer registry could achieve high accuracy for some cancer sites and may be practical and useful in the context of manual human review assistance. The cross validation and testing experiments showed that consolidation of M stage for the three cancer sites could achieve very high accuracy (93.9%–96.8%) while consolidation of T and N stages varied for different sites with the best performance observed for colon cancer cases (83.6%–91.2%), followed by prostate cancer cases (73.5%–81.4%) and lowest for lung cancer cases (60.4%–71.1%). Deriving a stage group from consolidated TNM stages on the testing subset showed high accuracy for colon cancer (88.4%) followed by lung cancer (84.5%) while accuracy for prostate cancer was lower (67.1%).

1.3 References

[1] David Blumenthal and Marilyn Tavenner. The meaningful use regulation for electronic health records. *N Engl J Med*, 2010(363):501–504, 2010.

[2] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.

[3] Carol Friedman, Thomas C Rindflesch, and Milton Corn. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765–773, 2013.

[4] F Martin-Sanchez and K Verspoor. Big data in medicine is driving big changes. *Yearbook of medical informatics*, 9(1):14, 2014.

[5] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[6] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[7] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform*, 35(128):44, 2008.

[8] Sumithra Velupillai, D Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of Medical Informatics*, 10(1):183, 2015.

[9] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[10] Marcelo Fiszman, Wendy W Chapman, Dominik Aronsky, R Scott Evans, and Peter J Haug. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604, 2000.

[11] Stephane Meystre and Peter J Haug. Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making*, 5(1):30, 2005.

[12] Imre Solti, Barry Aaronson, Grant Fletcher, Magdolna Solti, John H Gennari, Melissa Cooper, and Thomas Payne. Building an automated problem list based on natural language processing: Lessons learned in the early phase of development. In *AMIA Annual Symposium Proceedings*, volume 2008, page 687. American Medical Informatics Association, 2008.

[13] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[14] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[15] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.

[16] Chen Lin, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*, 22(e1):e151–e161, 2014.

[17] David S Carrell, Scott Halgrim, Diem-Thy Tran, Diana SM Buist, Jessica Chubak, Wendy W Chapman, and Guergana Savova. Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. *American Journal of Epidemiology*, 179(6):749–758, 2014.

[18] Yue Wang, Jin Luo, Shiying Hao, Haihua Xu, Andrew Young Shin, Bo Jin, Rui Liu, Xiaohong Deng, Lijuan Wang, Le Zheng, et al. Nlp based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. *International Journal of Medical Informatics*, 84(12):1039–1047, 2015.

[19] Riccardo Miotto and Chunhua Weng. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association*, 22(e1):e141–e150, 2015.

[20] Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014.

[21] Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

[22] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543, 2011.

[23] Son Doan, Mike Conway, Tu Minh Phuong, and Lucila Ohno-Machado. Natural language processing in biomedicine: A unified system architecture overview. *Clinical Bioinformatics*, pages 275–294, 2014.

[24] American Cancer Society. Cancer Facts & Figures 2017. `https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html`. Accessed: 12/05/2017.

[25] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: Cancer Journal for Clinicians*, 66(1):7–30, 2016.

[26] National Cancer Institute. Cancer statistics. `https://www.cancer.gov/about-cancer/understanding/statistics`. Accessed: 12/05/2017.

[27] BWKP Stewart, Christopher P Wild, et al. World cancer report 2014. *Health*, 2017.

[28] Donald M Parkin. The evolution of the population-based cancer registry. *Nature Reviews Cancer*, 6(8):603, 2006.

[29] Benjamin F Hankey, Lynn A Ries, and Brenda K Edwards. The surveillance, epidemiology, and end results program. *Cancer Epidemiology and Prevention Biomarkers*, 8(12):1117–1121, 1999.

[30] M Thornton and L OConnor. Standards for cancer registries volume II: Data standards and data dictionary, record layout version 12.2. *Springfield, IL: North American Association of Central Cancer Registries*, 2012.

[31] Calvin Zippin, Diana Lum, and Benjamin F Hankey. Completeness of hospital cancer case reporting from the seer program of the national cancer institute. *Cancer*, 76(11):2343–2350, 1995.

[32] National Cancer Registrars Association (NCRA). Become a cancer registrar. `http://www.ncra-usa.org/About/Become-a-Cancer-Registrars`. Accessed: 12/19/2017.

[33] David A Hanauer, Gretchen Miela, Arul M Chinnaiyan, Alfred E Chang, and Douglas W Blayney. The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. *Journal of the American College of Surgeons*, 205(5):690–697, 2007.

[34] Jan Willem Coebergh, Corina van den Hurk, Stefano Rosso, Harry Comber, Hans Storm, Roberto Zanetti, Lidia Sacchetto, Maryska Janssen-Heijnen, Melissa Thong, Sabine Siesling, et al. Eurocourse lessons learned from and for population-based cancer registries in europe and their programme owners: Improving performance by research programming for public health and clinical evaluation. *European Journal of Cancer*, 51(9):997–1017, 2015.

[35] R Zanetti, I Schmidtmann, L Sacchetto, F Binder-Foucard, A Bordoni, D Coza, S Ferretti, J Galceran, A Gavin, N Larranaga, et al. Completeness and timeliness: Cancer registries could/should improve their performance. *European Journal of Cancer*, 51(9):1091–1098, 2015.

[36] Stephen B Edge and Carolyn C Compton. The american joint committee on cancer: The 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of Surgical Oncology*, 17(6):1471–1474, 2010.

[37] Irena Spasić, Jacqueline Livsey, John A Keane, and Goran Nenadić. Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics*, 83(9):605–623, 2014.

[38] Abdulrahman Khalifa and Stéphane Meystre. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics*, 58:S128–S132, 2015.

[39] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *Journal of Biomedical Informatics*, 58:S67–S77, 2015.

[40] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[41] Abdulrahman Khalifa, Sumithra Velupillai, and Stephane Meystre. Utahbmi at semeval-2016 task 12: Extracting temporal information from clinical text. *Proceedings of SemEval*, pages 1256–1262, 2016.

[42] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062, 2016.

CHAPTER 2

ADAPTING EXISTING NATURAL LANGUAGE PROCESSING RESOURCES FOR RISK FACTORS IDENTIFICATION IN CLINICAL NOTES

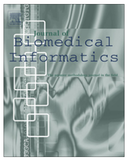
The article in this chapter was originally published in *Journal of Biomedical Informatics*, **58** (2015) S128–S132. It is reproduced here with permission of the publisher.

Journal of Biomedical Informatics 58 (2015) S128–S132

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes

Abdulrahman Khalifa *, Stéphane Meystre

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 17 February 2015
Revised 23 July 2015
Accepted 3 August 2015
Available online 28 August 2015

Keyword:
Natural language processing
Information extraction
Clinical narrative
Risk factors
Cardiovascular disease
Text mining
Medical records
Machine learning

ABSTRACT

The 2014 i2b2 natural language processing shared task focused on identifying cardiovascular risk factors such as high blood pressure, high cholesterol levels, obesity and smoking status among other factors found in health records of diabetic patients. In addition, the task involved detecting medications, and time information associated with the extracted data. This paper presents the development and evaluation of a natural language processing (NLP) application conceived for this i2b2 shared task. For increased efficiency, the application main components were adapted from two existing NLP tools implemented in the Apache UIMA framework: Textractor (for dictionary-based lookup) and cTAKES (for preprocessing and smoking status detection). The application achieved a final (micro-averaged) F_1-measure of 87.5% on the final evaluation test set. Our attempt was mostly based on existing tools adapted with minimal changes and allowed for satisfying performance with limited development efforts.

1. Introduction

The 2014 i2b2 (Informatics for Integrating Biology and the Bedside) challenge proposed several different tasks: clinical text de-identification, cardiovascular risk factors identification, software usability assessment, and novel data uses. Our efforts focused on the second track, identifying risk factors for heart disease based on the automated analysis of narrative clinical records of diabetic patients [1]. The annotation guidelines for the task defined eight categories of information associated with increased risk for heart disease: (1) Diabetes, (2) Coronary Artery Disease (CAD), (3) Hyperlipidemia, (4) Hypertension, (5) Obesity, (6) Family history of CAD, (7) Smoking and (8) Medications associated with the aforementioned chronic diseases. Each category of information (except family history of CAD and smoking status) had to be described with *indicator* and *time* attributes. The indicator attribute captures indications of the risk factor in the clinical text. For instance, Diabetes could be identified using a mention of the disease (i.e. "patient has h/o DMII"), or a hemoglobin A1c value above 6.5 mg/dL (i.e. "7/18: A1c: 7.3") while CAD could be identified using a mention (i.e. "PMH: significant for CAD"), or an event (i.e. "CABG in 1999"). The time attribute specifies the temporal relation to the Document Creation Time (DCT). It could take any one of the following values: before DCT, during DCT or after DCT. We refer the reader to [2] for a complete description of the annotation guidelines. For this challenge, we built a natural language processing (NLP) application based on the Apache UIMA (Unstructured Information Management Architecture) [3] and reusing existing tools previously developed to address similar tasks in previous i2b2 challenges. In this paper, we present our approach to extract relevant information from clinical notes, discuss performance results, and conclude with remarks about our experience adapting existing NLP tools.

2. Background

Extracting information from clinical notes has been the focus of a growing body of research these past years [4]. Common characteristics of narrative text used by physicians in electronic health records (e.g., telegraphic style, ambiguous abbreviations) make it difficult to access such information automatically. Natural Language Processing (NLP) techniques are needed to convert information from the unstructured text to a structured form readily processable by computers [5,6]. This structured information can then be used to extract meaning and enable Clinical Decision Support (CDS) systems that assist healthcare professionals and improve health outcomes [7]. Among the earliest attempts to develop NLP applications in the medical domain, the LSP (Linguistic String Project) [8], and MedLEE (Medical Language Extraction

* Corresponding author.
E-mail addresses: abdulrahman.aal@utah.edu (A. Khalifa), stephane.meystre@hsc.utah.edu (S. Meystre).

http://dx.doi.org/10.1016/j.jbi.2015.08.002
1532-0464/© 2015 Elsevier Inc. All rights reserved.

and Encoding system) [9] were prominent examples. More recent applications include MetaMap [10] developed by the National Library of Medicine to map terms in biomedical text with concepts in the UMLS (Unified Medical Language System) Metathesaurus [11]. cTAKES [12] was developed at the Mayo Clinic and is described as ''large-scale, comprehensive, modular, extensible, robust, open-source'' application based on Apache UIMA. It can be used to preprocess clinical text, find named entities and perform additional advanced NLP tasks such as coreference resolution. Textractor [13] is another UIMA-based application that was originally developed at the University of Utah to extract medications, their attributes, and reasons for their prescription from clinical notes.

When extracting information from clinical notes, NLP applications must take local contextual and temporal information into account for improved accuracy. Contextual information is important to determine if concepts are affirmed or negated (e.g., 'denies any chest pain'), or if the subject of the information is the patient or someone else (e.g., 'mother has diabetes'). Popular algorithms for negation detection in clinical notes include NegExpander [14] and NegEx [15]. Temporal information is critical to establish chronological order of events described in patient notes and to resolve mentions of procedures or laboratory results to specific time points for accurate analysis [16,17]. The ConText algorithm [18] proposed by Chapman et. al. is an extension of NegEx that allows analysis of contextual information like negation (negated, affirmed), temporality (historical, recent, hypothetical), and experiencer (patient, other). The development of NLP applications typically requires significant efforts and relies on annotated clinical text for training and testing. Widely accessible and shared annotated corpora in the medical domain are still rare, mainly because of strict patient privacy rules. This scarcity has been an obstacle to developing state-of-the-art NLP approaches for clinical text [19]. To address this obstacle and enable direct comparison of NLP approaches in the clinical domain, i2b2 shared NLP tasks have been organized almost annually since 2006. The challenges started with an automated de-identification [20] and smoking status detection [21] challenges. In 2008, the i2b2 challenge focused on identifying information about obesity and 15 co-morbidities [22]. In 2009, the third i2b2 challenge [23] was focused on identifying medications and associated information such as dosage and frequency. This was followed by challenges for medical concept extraction, assertion and relations classification in 2010 [24], followed by coreference resolution tasks in 2011 [25] and a temporal relations classification in 2012 [26].

To reduce development efforts, many authors have reused NLP tools or resources such as ConText, sentence boundary detectors and part-of-speech taggers from OpenNLP project [27], the Stanford parser [28], or the Weka machine learning framework [29], but the majority of their applications were still new developments. Reusing larger components or even existing NLP applications could allow for further development effort reduction. A good example was the application developed by Wellner et al. [30] for the 2006 i2b2 de-identification task. It was based on the adaptation of two applications originally designed for recognizing named entities in newswire text. The process involved running two applications out-of-the-box as a baseline and then gradually introducing a few task-specific features, using bias parameters to control feature weights, and adding lists of common English words during development to improve performance. With minimal effort, they were able to obtain very high performance for the task. Although their attempt used applications out-of-the-box as baselines, they had to re-train the models with new task-specific features to achieve high performance. Our attempt focused on adapting existing tools that were developed to solve similar tasks in the past, and do it without feature engineering and re-training of machine learning models.

3. Methods

3.1. Datasets

The i2b2 NLP shared task organizers distributed two annotated datasets (SET1 and SET2) to be used for development and training. These sets were released separately, with a few weeks interval. SET1 was composed of 521 de-identified clinical notes and SET2 was composed of 269 de-identified notes; therefore, a total of 790 documents were available for training. The test set was released three days before final submission and consisted of a total of 514 de-identified clinical notes.

3.2. NLP application overview

As already mentioned, our application was based on the Apache UIMA framework, with components adapted from two existing applications. Because of the various nature of information to be extracted in this task, we experimented with different approaches for different categories of information. For example, Textractor's dictionary-based lookup component was used to detect mentions of chronic diseases, in addition to mentions of CAD events as defined in the annotation guidelines. The results of the lookup module were then filtered using lists of UMLS Metathesaurus concept identifiers CUIs for disease and risk factor concepts defined for the task. Smoking status was identified using the existing classifier available from cTAKES. Medications and the various test results (hemoglobin A1c, glucose, blood pressure, cholesterol, etc) were identified using pattern matching with regular expressions. Family history of CAD was detected by modifying the contextual analysis of the detected CAD mentions using ConText's 'experiencer' analysis.

The application pipeline is depicted in Fig. 1 and described below. The analysis of clinical text begins with a preprocessing stage that consists in segmenting the text into sections, splitting it into sentences, tokenizing and assigning part-of-speech tags to the input text with cTAKES. This is followed by running the smoking status classifier from cTAKES ''out-of-box'' to classify each patient record to a smoking status category: CURRENT, PAST, EVER, NEVER, UNKNOWN. The existing cTAKES SMOKER label was changed to EVER, as defined for this i2b2 task.

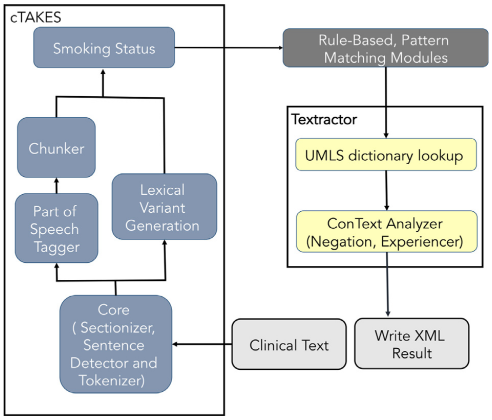

cTAKES
Smoking Status
Chunker
Lexical Variant Generation
Part of Speech Tagger
Core (Sectionizer, Sentence Detector and Tokenizer)
Rule-Based, Pattern Matching Modules
Textractor
UMLS dictionary lookup
ConText Analyzer (Negation, Experiencer)
Clinical Text
Write XML Result

Fig. 1. Overview of NLP application pipeline with adapted components from cTAKES and Textractor.

The text analysis then continues with rule-based pattern matching modules for detecting medications and laboratory test results. Medications were detected with a manually curated terminology of synonymous terms and abbreviations linked to each medications category. These lists were compiled using UMLS Metathesaurus terminologies and lists of common abbreviations found in clinical narratives (manually built by local domain experts); and then manually grouping the concepts into medication categories. The number of terms used for each medications varied widely, ranging from as few as 3 (e.g. for metformin) to more than 50 (e.g. for beta blockers and aspirin). Laboratory test results and vital signs were detected using regular expressions and the associated values were compared with abnormality thresholds defined in the guidelines. For instance, the phrase "Cholesterol-LDL 08/26/2091 148" indicates an LDL cholesterol concentration of 148 mg/dL, which is above the normal concentration of 100 mg/dL and should therefore be included as a risk factor. Special attention was paid to avoid incorrect values that were part of other numeric expressions (e.g., dates) by restricting regular expression matches to reasonable value ranges and imposing specific conditions on number boundaries (see examples in Table 1). Two regular expressions were used for each relevant laboratory test or vital sign indicator; one for capturing the term and the other for numerical value associated with the laboratory test or vital sign.

The application then proceeded with the UMLS Metathesaurus lookup module from Textractor. This module uses Apache Lucene-based [31] dictionary indexes to detect disease and risk factor terms. Before the dictionary lookup, acronyms were expanded and tokens normalized by removing unwanted stopwords. The lookup module then matched terms that belonged to one of the predefined UMLS semantic types for diseases (i.e., T019, T033, T046, T047 and T061). Matching was performed at the token level first, and then expanded to match at the noun phrase chunk level. All detected concepts were then filtered based on their CUIs to only include concepts belonging to one of the five disease and risk factor categories identified in the guidelines: CAD, Diabetes mellitus, Obesity, Hyperlipidemia, and Hypertension.

Finally, the application performed contextual analysis of all extracted and filtered information to exclude negated concepts, verify that the patient was the experiencer, and produce time attributes for each concept in relation to the DCT. Negation and experiencer analysis was performed using a local implementation of the ConText algorithm, as available in Textractor. Detection of family history of CAD was handled by considering all extracted CAD concepts with an experiencer other than the patient (e.g., "mother has history of CAD") as a *present* family history of CAD. If all CAD concepts were identified as belonging to the patient, or if no CAD concepts were found in the clinical note, then family history of CAD was set to *not present*.

We experimented with various uses of ConText's temporal analysis (i.e., concepts classified as recent, historical or hypothetical) in order to map them to the corresponding time attribute values (i.e., before DCT, during DCT or after DCT). However, initial results on the training data using this approach were not satisfying. As an alternative approach, we used the most common time value found for each category of information in the training data. For example, chronic diseases such as CAD and most medications were mostly *continuing* (i.e., existed before, during, and after the hospital stay or visit) and therefore annotated with all three time attribute values in the reference standard. As another example, laboratory test results varied with examples like hemoglobin A1c and glucose tests that were mostly 'before DCT', and others like hypertension that were mostly 'during DCT'.

4. Results

After development and refinement based on the training corpus (SET1 and SET2), the NLP application processed the testing corpus when made available, and the application output was sent to the shared task organizers for analysis. The application output was compared with the reference standard using the evaluation script provided by the shared task organizers and all extracted information classified as true positive (i.e., output matches with the reference standard), false positive, or false negative. Metrics used included recall, precision, and the F_1-measure (details in [1]). The results for each class of information are presented in Table 2. For overall averages, both macro- and micro-averages are included. Each separate class-indicator combination is reported using micro-averages only. The evaluation script contained an option to calculate results separately for each class of information using the `--filter` option. It also allowed computing specific class and indicator attribute values such as the class DIABETES and indicator attribute value of *mention* using the option `--conjunctive`. Results for each disease category are presented for *mention* and each disease-specific indicators separately as in the annotation guideline. The SMOKING category results are presented as *status*

Table 2
Macro- and micro-averaged overall results including the micro-averaged breakdown of final results for every class of information given in terms of Precision, Recall and F_1-measure.

	Indicator	Precision	Recall	F_1-measure
CAD	Mention	0.883	0.9651	0.9222
	Symptom	0.2095	0.4429	0.2844
	Event	0.6457	0.5899	0.6165
	Test	0.4557	0.6102	0.5217
Diabetes	Mention	0.9512	0.9887	0.9696
	A1C	0.8611	0.7561	0.8052
	Glucose	0.1486	0.3333	0.2056
Hyperlipidemia	Mention	0.9899	0.827	0.9011
	High cholesterol	0.5714	0.3636	0.4444
	High LDL	0.84	0.7241	0.7778
Hypertension	Mention	0.9918	0.9891	0.9904
	High BP	0.8571	0.5231	0.6497
Obesity	Mention	0.7562	1.0	0.8612
	BMI	0.9231	0.7059	0.8
Smoking		0.8638	0.8672	0.8655
Medication		0.8282	0.8911	0.8585
Family history of CAD		0.9494	0.9494	0.9494
Macro-average		0.8494	0.8914	0.8699
Micro-average		0.8552	0.8951	0.8747

Table 1
Examples of regular expressions used for matching test mentions and values.

<table>
<tr><th>Laboratory/test</th><th>Regular expression for mention</th><th>Regular expression for value</th></tr>
<tr><td>Glucose (for Diabetes mellitus)</td><td><code>(fasting)? (blood)? (glucose|\bGLU(-poc)?\b|\bBG\b|(blood) sugar(s)?|\bFS\b|\bBS\b|fingerstick|\bFG\b)</code></td><td><code>(?<!/|\d)(\d\d\d?)(-\d\d\d)?(?!/|\d|\w)</code></td></tr>
<tr><td>Blood pressure (for Hypertension)</td><td><code>(?<!\w)((s)?BP[s]?|b/p|((blood|systolic)[]+pressure[s]?)|hypertensive)[:]? (?!\w)</code></td><td><code>(?<!/|\d)(\d\d\d)/(\d\d\d?) (?!/\d|\d)</code></td></tr>
</table>

only, and MEDICATION results are aggregated for all the categories correctly identified in the clinical records. All results in the table were computed for all three values of time attribute for each class and no attempt made to separate 'before DCT', 'during DCT' and 'after DCT' results for each class.

As shown in Table 2, the application achieved an overall micro-averaged F_1-measure of 87.47% and a macro-averaged F_1-measure of 86.99%. In most disease categories, accuracy was highest for mentions of disease with micro-averaged F1-measures of 92.22%, 94.94%, 96.96%, 90.11%, and 99.04% for CAD, family history of CAD, Diabetes, Hyperlipidemia, and Hypertension, respectively. Medications, mentions of Obesity and Smoking status identification accuracy reached micro-averaged F_1-measures of 85.85%, 86.12% and 86.55%, respectively. Accuracy was lower with other information categories such as laboratory tests, CAD events and symptoms with F_1-measures ranging from 20.56% to 80%.

5. Discussion

As presented above, the application accuracy for mentions of the various diseases, smoking status, medications and family history was higher than accuracy for any other indicator type defined in the annotation guidelines (e.g., laboratory tests, CAD events and symptoms). The dictionary lookup approach with terminological content from the UMLS Metathesaurus for detecting disease mentions was successful for this task. Similarly, the smoking status classifier from cTAKES successfully identified and classified smoking status information (F_1-measure of about 87%) despite the fact that the model was used out-of-the-box, without any training on the new corpus for the current i2b2 NLP task. The identification of medications and their attributes reached an F_1-measure of about 86% when using regular expressions and manually curated lists of terms, demonstrating the feasibility of this approach for the type of narrative notes used in this shared task. The precision obtained for medications was lower (83%) than recall (89%) and hence affected the final F_1-measure. This is mainly due to the way we chose to generate the time attribute by using the continuing times scenario (i.e., generating 'before DCT', 'during DCT' and 'after DCT' temporal information tags for every medication detected in the notes). Obviously, there will be false positives associated with this approach when medications strictly occur for either one or two of the time values in the clinical notes. In addition, since the medication term lists were created manually, some spelling variations and terms could have been missed, therefore producing some false negatives and affecting overall recall. An example of spelling variation is the term 'nitroglycerine' in the *nitrate* group category, which appeared in both corpora as 'nitroglycerin'. The latter was not in the nitrate list used by our application and hence caused some false negatives. An example of completely missed terms was sublingual nitroglycerin mentioned as 'SL NTG'. Among disease mentions, the Hyperlipidemia class had the lowest recall (83%) and Obesity had the lowest precision (76%). The former was mostly due to some clinical reports containing annotations for Hyperlipidemia mentions appearing as 'elevated serum cholesterol', 'elevated lipids' and 'high cholesterol' that were missed by our application because of inaccurate chunking. In addition, we did not have the corresponding CUI codes for some of them in our dictionary lookup module. There were at least two cases in the testing corpus where Hyperlipidemia was mentioned directly following a word with no space in between such as 'hemodialysisHyperlipidemia' which our application missed also. The low precision with Obesity was caused by including the UMLS concept 'overweight' in our list of CUIs for Obesity. Although 'overweight' was used as indicator for obesity in one record in the reference standard corpora, its use produced many false positives since 'overweight' often does not indicate obesity. There were also false positive mentions of Obesity produced by our application in cases where 'obese' was mentioned without indicating Obesity (e.g., "abdomen is slightly obese" and "Abdomen: Moderately obese"). The other indicators for diseases and risk factors were quite challenging and our approach using regular expressions at the lexical level was not always effective. With the exception of hemoglobin A1c laboratory tests (for Diabetes), BMI (for Obesity), and cholesterol LDL (for Hyperlipidemia), the application performance was modest with an F_1-measure ranging from 21% for the blood glucose indicator up to 65% for the blood pressure indicator. Some of the challenges with these indicators are summarized below:

- **Lexical and spelling variations:** Some laboratory indicators for diseases are mentioned with many lexical variations and acronyms. Table 1 shows the regular expressions used to capture blood glucose for diabetes and blood pressure for hypertension. As shown, glucose can be described with a variety of terms like BG, BS, FS and FG; and blood pressure can be described with terms like BP and b/p. This is an example of some of the limitations with our approach. and a comprehensive strategy to deal with this issue to enable better accuracy would be needed.
- **Extracting laboratory numerical results accurately:** When the application finds matching terms for laboratory or test indicators, it must proceed with extracting associated numerical values and compare them to threshold levels for abnormality. Extracting numerical values may be straightforward when they immediately follow the term and are expressed as single units such as in the phrase "FSBG was 353". However, other phrases can be more challenging like "FG 120–199; now 68–172, although 172 = outlier, mostly in the 70–130". In this case, ranges of values are expressed with '—', and multiple units are expressed with temporal and frequency modifiers (i.e. 'now' and 'mostly').
- **Training data sparseness:** The number of training examples available was sometimes too low to allow for the variety needed for adequate application generalization. For instance, in the case of cholesterol indicator for Hyperlipidemia, the total number of available annotations was only 9 in the whole set of 790 training documents. In contrast, there were about 33 annotations available for the LDL indicator.
- **Complex time analysis**. Test and laboratory indicators require more sophisticated time attribute analysis and this is another limitation of our approach. Unlike chronic disease mention annotations which were mostly characterized with 'continuing' time attribute (i.e. before, during and after DCT), most of the laboratory and vital sign annotations were characterized by a variety of time attribute values. For instance, hemoglobin A1c and glucose tests were usually conducted in a prior visit and hence mostly annotated with 'before DCT' while blood pressure (BP) was mostly measured during the patient visit and hence had mostly 'during DCT' time value. To examine the impact of time attributes on performance of our application, we followed the "fixed" evaluation procedure described in [32] and produced results for some indicators after replacing the value of time attribute with 'before DCT' in all annotations from our application output and in the testing reference standard (see Table 3). This evaluation considers true positives, false positives and false negatives for each individual annotation while ignoring the time attributes (i.e. application output is not penalized for incorrect time values). As shown in Table 3, the performance of our application improved when the time component was ignored in the evaluation (compare with results from Table 2). Our decision to use the most common time attribute values for each of these indicators caused a loss in precision and recall contributing to lower overall F_1-measure score.

Table 3
Results for Medications and some disease indicators after fixing the time attribute to the same value in both application output and testing reference standard.

	Precision	Recall	F$_1$-measure
Glucose	0.2568	0.6129	0.3619
High cholesterol	0.7143	0.5	0.5882
High BP	0.8908	0.5792	0.702
MEDICATIONS	0.8791	0.8826	0.8808

6. Conclusion

Our rapid approach, adapting resources from existing applications for the 2014 i2b2 challenge, allowed for performance similar to other more sophisticated application developed for this task which used additional manual annotations or multiple machine learning classifiers [1]. We think that existing NLP resources should be reused, and most can be adapted and used at least as baseline for future tasks in the clinical domain. Improvements for future attempts shall focus on a comprehensive strategy to tackle spelling errors and variations, acronyms disambiguation, and more refined temporal analysis. Use of standard terminologies, as available in the UMLS Metathesaurus, should be the basis for these clinical information extraction tasks as they already contain well-defined concepts associated with multiple terms. Finally, regular expressions and pattern matching can be useful for extracting information such as name-value pairs from short phrases (e.g. 'Cholesterol- LDL 08/26/2091 148'). However, longer phrases containing complex syntactic structures require the use of advanced parsing techniques to identify constituents and relations between them. In the future, we plan to explore advanced techniques such as dependency parsing or semantic role labeling to reduce errors appearing with long phrases requiring deeper contextual analysis to be accurately extracted. For instance, in the following sentence: "Prior to her bypass surgery on the right leg, she underwent a Persantine MIBI which showed only 1 mm ST depressions and was considered not diagnostic"; it is important for an application to link the negated phrase "was considered not diagotstic" with the noun phrase "Persantine MIBI" to conclude that although the patient had the MIBI test performed, the result was not diagnostic and therefore the test indicator (i.e. 'MIBI') ruled out CAD.

Conflict of interest

Authors do not have conflict of interest.

References

[1] A. Stubbs, C. Kotfila, H. Xu, Ö. Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform. 58S (2015) S67–S77.

[2] A. Stubbs, Ö. Uzuner, Annotating risk factors for heart disease in clinical narratives for diabetic patients, J. Biomed. Inform. 58S (2015) S78–S91.

[3] D. Ferrucci, A. Lally, Uima: an architectural approach to unstructured information processing in the corporate research environment, Nat. Language Eng. 10 (3–4) (2004) 327–348.

[4] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, et al., Extracting information from textual documents in the electronic health record: a review of recent research, Yearb. Med. Inform. 35 (2008) 128–144.

[5] A. Pratt, Medicine computers and linguistics, Biomed. Eng. (1973) 87–140.

[6] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (5) (2011) 544–551.

[7] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can natural language processing do for clinical decision support?, J Biomed. Inform. 42 (5) (2009) 760–772.

[8] E. Chi, M. Lyman, N. Sager, C. Friedman, C. Macleod, A database of computer-structured narrative: methods of computing complex relations, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1985, p. 221.

[9] C. Friedman, S.B. Johnson, B. Forman, J. Starren, Architectural requirements for a multipurpose natural language processor in the clinical environment, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1995, p. 347.

[10] A.R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[11] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucl. Acids Res. 32 (Suppl. 1) (2004) D267–D270.

[12] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.

[13] S.M. Meystre, J. Thibault, S. Shen, J.F. Hurdle, B.R. South, Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents, J. Am. Med. Inform. Assoc. 17 (5) (2010) 559–562.

[14] D.B. Aronow, F. Fangfang, W.B. Croft, Ad hoc classification of radiology reports, J. Am. Med. Inform. Assoc. 6 (5) (1999) 393–411.

[15] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310.

[16] L. Zhou, G.B. Melton, S. Parsons, G. Hripcsak, A temporal constraint structure for extracting temporal information from clinical narrative, J. Biomed. Inform. 39 (4) (2006) 424–439.

[17] P. Bramsen, P. Deshpande, Y.K. Lee, R. Barzilay, Finding temporal order in discharge summaries, AMIA Annual Symposium Proceedings, vol. 20, American Medical Informatics Association, 2006, p. 81.

[18] W.W. Chapman, D. Chu, J.N. Dowling, ConText: an algorithm for identifying contextual features from clinical text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, 2007, pp. 81–88.

[19] W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, Ö. Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, J. Am. Med. Inform. Assoc. 18 (5) (2011) 540–543.

[20] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.

[21] Ö. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inform. Assoc. 15 (1) (2008) 14–24.

[22] Ö. Uzuner, Second i2b2 workshop on natural language processing challenges for clinical records, in: AMIA. Annual Symposium proceedings/AMIA Symposium, AMIA Symposium, 2007, pp. 1252–1253.

[23] Ö. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (5) (2010) 514–518.

[24] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.

[25] Ö. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc., 2012 (amiajnl–2011).

[26] W. Sun, A. Rumshisky, Ö. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 challenge, J. Am. Med. Inform. Assoc., 2013 (amiajnl–2013).

[27] T. Morton, J. Kottmann, J. Baldridge, G. Bierner, Opennlp: A java-based nlp toolkit, 2005.

[28] M.-C. De Marneffe, B. MacCartney, C.D. Manning, et al., Generating typed dependency parses from phrase structure parses, in: Proceedings of LREC, vol. 6, 2006, pp. 449–454.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.

[30] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, J. Hitzeman, L. Hirschman, Rapidly retargetable approaches to de-identification in medical records, J. Am. Med. Inform. Assoc. 14 (5) (2007) 564–573.

[31] A. Białecki, R. Muir, G. Ingersoll, Apache lucene 4, in: SIGIR 2012 Workshop on Open Source Information Retrieval, 2012, pp. 17–24.

[32] C. Grouin, V. Moriceau, P. Zweigenbaum, Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records, J. Biomed. Inform. 58S (2015) S133–S142.

CHAPTER 3

UTAHBMI AT SEMEVAL-2016: EXTRACTING TEMPORAL INFORMATION FROM CLINICAL TEXT

The article in this chapter was originally published in *Proceedings of SemEval-2016*, (2016) 1256–1262. It is reproduced here with permission of the publisher.

UtahBMI at SemEval-2016 Task 12: Extracting Temporal Information from Clinical Text

Abdulrahman Khalifa
University of Utah
abdulrahman.aal@utah.edu

Sumithra Velupillai
KTH, Stockholm/King's College, London
sumithra@kth.se

Stephane Meystre
University of Utah
stephane.meystre@hsc.utah.edu

Abstract

The 2016 Clinical TempEval continued the 2015 shared task on temporal information extraction with a new evaluation test set. Our team, UtahBMI, participated in all subtasks using machine learning approaches with ClearTK (LIBLINEAR), CRF++ and CRFsuite packages. Our experiments show that CRF-based classifiers yield, in general, higher recall for multi-word spans, while SVM-based classifiers are better at predicting correct attributes of TIMEX3. In addition, we show that an ensemble-based approach for TIMEX3 could yield improved results. Our team achieved competitive results in each subtask with an F1 75.4% for TIMEX3, F1 89.2% for EVENT, F1 84.4% for event relations with document time (DocTimeRel), and F1 51.1% for narrative container (CONTAINS) relations.

1 Introduction

Extracting temporal information from unstructured clinical narratives is an important step towards the accurate construction of a patient timeline over the course of clinical care (Savova et al., 2009), identifying and tracking patterns of care that are crucial for decision making (Augusto, 2005; Wang et al., 2008) and identifying cases or cohorts with temporal criteria for medical research (Raghavan et al., 2014). In the medical domain, more emphasis has been placed on utilizing temporal information from structured databases (Combi et al., 2010). However, recent developments in Medical Natural Language Processing (NLP) research has stimulated work in extracting information from unstructured clinical text (Meystre et al., 2008; Velupillai et al., 2015a) and facilitated future directions to extracting temporal information (Zhou and Hripcsak, 2007).

The i2b2 series of NLP challenges focused in 2012 on extracting events (problems, treatments and tests), time expressions (date, duration, time and frequency) and temporal relations (before, after, overlap) from a set of annotated discharge summaries. The best performing systems used supervised machine learning approaches, except for time expression identification and normalization where rule-based followed by hybrid approaches were most successful (Sun et al., 2013b; Sun et al., 2013a).

In 2015, the SemEval challenge included a Clinical TempEval task (Bethard et al., 2015) with similar objectives to the 2012 i2b2 challenge. The TimeML event and temporal expressions specification language (Pustejovsky et al., 2010) was adapted to define events, time expressions and relation annotations suitable for the clinical domain (Styler et al., 2014). The THYME (Temporal Histories of Your Medical Event) corpus is used in the Clinical TempEval challenge. The annotations in this corpus introduce the use of narrative containers concept (Pustejovsky and Stubbs, 2011) to reduce the complexity of finding temporal relations between every possible pair, and allow rapid discovery through automatic inferences. Each event and time expression is, when possible, assigned a narrative container that defines their temporal span. Groups of events and times within a narrative container can then be linked as one unit with other containers; eliminating the need to explicitly link every pair of events and times.

Proceedings of SemEval-2016, pages 1256–1262,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

The additional pairs can be derived easily from minimal links between pairs within different narrative containers.

We present in this paper the methods used and results obtained from experiments with SVM-based linear classifiers and CRF-based sequential classifiers for the Clinical TempEval task. We complement the paper with a discussion and insights that potentially could help future efforts in this domain.

2 Methods

2.1 Task & Materials

The 2016 Clinical TempEval challenge included 6 subtasks: TIMEX3 1) span detection and 2) attribute classification, EVENT 3) span detection and 4) attribute classification, 5) relation between each event and document creation time classification (known as DocTimeRel), and narrative container or 6) CONTAINS relations between pairs of events and times classification. Our team participated in both phases provided in the challenge (phase 1: plain text only of the test set and phase 2: reference annotations for TIMEX3 and EVENTS including attributes were given for the relation classification subtasks) For a detailed description of the subtasks and evaluation metrics we refer the reader to (Bethard et al., 2015; Bethard et al., 2016).

The THYME corpus used in this task consists of treatment and pathology notes for colon cancer patients from the Mayo clinic. Three datasets were provided: *train* (=293 documents), *dev* (=147) and *test* (=151). We used the dev set to benchmark different approaches during system development and as a guideline to manually select the best performing features. All final models used for predictions were trained using the combined *train*+*dev* datasets. The test set was used for the final evaluation. Each subtask was addressed separately using a machine learning classifier and groups of almost similar features with slight changes such as surrounding context window sizes. cTAKES (Savova et al., 2010) was used to pre-process each clinical note to generate morphological, lexical and syntactic-level annotations, which were used as features for training the classifiers. The ClearTK machine learning package (Bethard et al., 2014) was used to build Support Vector Machine (SVM) LIBLINEAR (Fan et al., 2008) classifiers, while CRFsuite (Okazaki, 2007) and CRF++ (Kudo, 2005) were used to build Conditional Random Field (CRF) sequential classifiers. Both cTAKES and ClearTK utilize the Apache Unstructured Information Management Applications (UIMA) framework (Ferrucci and Lally, 2004) which makes it easy to integrate modules from both applications and pipeline output from cTAKES to ClearTK using the XML Metadata Interchange (XMI) format.

2.2 Input Preparation/Feature Extraction

Each clinical note in the corpus was previously segmented into sections with a `[start section id=...]` and `[end section id=...]` markers that were easy to identify and annotate using regular expressions. Therefore, we built a UIMA module to segment each clinical note into section boundaries; each annotated with their respective section ID. cTAKES clinical pipeline (version 3.2.2) was used to extract lexical and syntactic features. These include sentence boundaries, tokens, lemmas, part-of-speech tags, syntactic chunk tags (e.g. Verb Phrase-VP, Noun Phrase-NP), token type as defined by cTAKES (see figure 1), as well as dependency parse and semantic role labels used for relation classification. Furthermore, ClearTK feature extractors were used to generate word shape features (e.g. capital, lower, numeric), character patterns and character N-gram features for the linear classifiers. The CRFsuite package comes with built-in feature extractor functions for word shapes, character patterns and N-gram which were used for the TIMEX3, EVENT and DocTimeRel CRF classifiers. Table 1 outlines the features used in each subtask.

For the CRF packages, the features had to be transformed into a flat, tab-separated structure with columns of tokens and associated features each placed in one line. Sentences are designated by empty lines following a sequence of lines of tokens (see Figure 1 for an example).

2.3 SVM-based Approach

The LIBLINEAR package within ClearTK was used to train all linear classifiers with default settings (C=1.0; s=1; Loss=dual L2-regularized) except for TIMEX3 (grid search performed on the training set indicated a better value for C=0.5). We re-used

	CRFsuite			CRF++	LIBLINEAR			
Feature Type	TIMEX3	EVENT	DocTimeRel	DocTimeRel	TIMEX3	EVENT	DocTimeRel	CONTAINS
Window Size (preceding, following)	$-2, +2$	$-2, +2$	$-2, +2$	$-5, +5$	$-5, +5$	$-2, +2$	$-5, +5$	$-5, +5$
Token	*	*	*	*	*	*	*	*
Token (lowercased)	*	*	*		*	*		
Lemma	*	*	*	*			*	
Part of Speech (POS)	*	*	*	*	*	*	*	*
Chunk Type	*	*	*	*				
Token Type (WORD, NUMERIC, ...)	*	*	*	*	*			
Word Shape (ALL-CAP, INITIAL-CAP, ...)	*	*	*		*			
Section ID	*	*	*	*	*	*	*	*
Character Pattern	*	*	*		*			
Character Ngram	*	*	*		*			
EVENT and attributes Tags							*	*
TIMEX3 and attributes Tags							*	*
HeidelTime Token					*			
TIMEX position in sentence								*
Number of tokens between relation pair								*
Semantic role arguments								*
C Parameter	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.0

Table 1: List of features used (indicated with asterisk) for each subtask with different machine learning approaches.

```
token   lemma   pos chunk   token_type      section_ID  ...
#       #       NN  B-NP    SymbolToken     20112       ...
1       1       LS  I-NP    NumToken        20112       ...
Dilated dilat   JJ  I-NP    WordToken       20112       ...
```

Figure 1: Example of the flat input used for the CRF approaches: features in columns separated by tabs.

the approach taken in the 2015 Clinical TempEval (Velupillai et al., 2015c) for TIMEX3, EVENT and DocTimeRel subtasks, with minor changes in the used features. For TIMEX3, one separate classifier was created for each class (e.g., DATE, TIME). For EVENT, one classifier was created for detecting the text span, and one separate classifier for each attribute (i.e., MODALITY, DEGREE, POLARITY and TYPE). In addition, we added a classifier in this pipeline, for event relations with the document time (DocTimeRel). The main feature additions in this year's challenge were a section ID feature for all classifiers; and a binary feature— whether or not a token was classified as temporal expression of an adapted version of HeidelTime (Strötgen and Gertz, 2010) — for the TIMEX3 subtask.

For the narrative container (CONTAINS) relations subtask, we trained four models to predict relations between pairs of 1) event-event and 2) event-time within a sentence; and 3) event-event and 4) event-time across consecutive sentences. This approach has been previously shown to be most effective in predicting temporal relations (Xu et al., 2013). The candidate pairs were selected[1] using the following strategy: All possible combinations of pairs between events and events-times within a sentence were generated for training and classification. For event-event pairs across consecutive sentences; only the first and last event from the current sentence were paired with the first and last from next (or subsequent) sentence. For event-time pairs across sentences; each time phrase in the current sentence is paired with the first and last events from the preceding and following sentences. This approach suffers from the limitation of allowing many examples with the negative class (i.e., pairs without a relation) to be selected; and hence causes class imbalance that may affect classifier training. (Tang et al., 2013) demonstrated that using heuristics to select candidates that are more likely to be part of a relation could produce superior results for temporal relation classification. Another possible remedy is to introduce scaling parameters to adjust the weight of each class during training, such that data samples from the positive class get more weight while the negative class samples get less weight (Lin et al., 2015). Due to time constraints, we were unable to experiment with either of these approaches.

2.4 CRF-based Approach

For the sequential classification, we used the CRFsuite for TIMEX3, EVENT and DocTimeRel subtasks in phase 1, and CRF++ for the DocTimeRel subtask in phase 2. All CRF trained models used default settings (C=1.0; algorithm=L-BFGS). During phase 1, we employed a cascaded approach: we trained CRFsuite models to 1) predict textual spans of TIMEX3 and EVENT tokens separately; 2) pre-

[1] cTAKES Temporal module was very useful in facilitating experiments for the TLINK relations.

	span			span+class		
	P	R	F1	P	R	F1
MAX	0.840	0.758	0.795	0.815	0.735	0.772
CRFsuite	0.798	0.714	0.754	0.771	0.690	0.729
LIBLINEAR	**0.810**	0.690	0.745	**0.792**	0.674	0.728
CRFsuite+LIBLINEAR	0.761	**0.769**	**0.765**	0.733	**0.741**	**0.737**
memorize (Baseline)	0.774	0.428	0.551	0.746	0.413	0.532

Table 2: TIMEX3 subtask results on the test set.

dict TIMEX3 and EVENT attributes using the predictions in step 1), and 3) predict DocTimeRel and CONTAINS relations using the predictions in steps 1-2. The prediction labels were encoded using the standard IOB2 format of **I**nside, **B**egin, and **O**utside. For instance, prediction labels for the phrase "see him this afternoon ." will be encoded as "O O B-TIME I-TIME O" where "this afternoon" is a TIMEX3 expression in this context. CRF classifiers are probabilistic graphical models that take into account a previous window of prediction labels and assign the most likely sequence of labels based on estimates obtained from the training data. Therefore, they usually perform better in tasks that require assigning labels to sequential data. This is particularly true for the TIMEX3 subtask where the majority of time phrases span multiple tokens.

3 Results

The performance we obtained for the various subtasks on the test set are shown in Tables 2, 3, 4. We also include the results from two baseline systems (**memorize** — for EVENT, TIMEX3 and DocTimeRel, and **closest** — for CONTAINS relations) provided by the workshop organizers, as well as the maximum score achieved in each subtask from all submissions (Bethard et al., 2016). Note that for the narrative container subtask, we report the official score and corrected score we obtained after discovering and correcting a bug affecting the LIBLINEAR models that prevented predictions of event-time relations.

CRF achieved a better performance (F1 %75.4) than the linear classifier (F1 %74.5) when detecting TIMEX3 spans because of higher recall (R %71.4). The LIBLINEAR model resulted in higher precision (P %81). Our initial analysis indicates that this is partly due to many CRF predictions overlapping with the reference annotations rather than matching exactly. When using a strict match evaluation approach, these overlaps are counted as false positives. For example, the CRF approach generated TIMEX3 labels for expressions like "at the time" and "in the past" while the reference standard included TIMEX3 annotations for only "the time" and "past", respectively. Combining the predictions from both models (by taking the union set of outputs and discarding duplicated predictions) allowed for improved performance (F1 %76.5) suggesting that an ensemble-based strategy could yield superior results for this subtask. Additional analysis will be needed to understand which class of TIMEX3 phrases each model is better at predicting and apply a more sophisticated ensemble method such as weighted average.

The results for the EVENT subtasks were almost identical between the two approaches (CRF or LIBLINEAR), except when classifying the *modality* and *type* attributes where CRF performed better. Combining the predictions from both models did not allow for any performance improvements. Note also that the baseline results for this subtask are very high.

For the DocTimeRel subtask, the CRFsuite model reached an F1 of %74.5 in phase 1, while the CRF++ model reached an F1 of %84.4 in phase 2; allowing for significant improvement over the performance of the LIBLINEAR model (F1 %81.8). For the CONTAINS relations classification subtask, the LIBLINEAR models achieved an F1 of %42.2 in phase 1 when using CRF predictions of TIMEX3 and EVENT; and F1 of %51.1 in phase 2. Note that for phase 2 we also included the prediction of DocTimeRel relations from CRF as an input feature to the LIBLINEAR models.

4 Discussion

Several important issues need to be addressed for future improvement in this task or other similar tasks. We outline some of these issues below, along with an analysis from the reference standard annotations and the system prediction errors.

The CRF-based classifiers detected TIMEX3 mentions with higher accuracy. As mentioned previously, many of these mentions were overlapping with the reference standard annotations. Our output included 352 false positive errors when using a strict match evaluation. Among these errors, about

	span			span+modality			span+degree			span+polarity			span+type		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MAX	0.915	0.891	0.903	0.866	0.843	0.855	0.911	0.887	0.899	0.900	0.875	0.887	0.894	0.870	0.882
CRFsuite	0.902	0.883	0.892	0.850	0.832	0.841	0.898	0.879	0.889	0.885	0.867	0.876	0.875	0.857	0.866
LIBLINEAR	0.897	0.886	0.892	0.841	0.831	0.836	0.892	0.881	0.887	0.879	0.869	0.874	0.854	0.843	0.849
memorize (Baseline)	0.878	0.834	0.855	0.810	0.770	0.789	0.874	0.831	0.852	0.812	0.772	0.792	0.855	0.813	0.833

Table 3: EVENT subtask results on the test set.

	DocTimeRel			CONTAINS		
	P	R	F1	P	R	F1
Phase 1: End-to-End with plain text only						
MAX	0.766	0.746	0.756	0.531	0.471	0.479
CRFsuite	0.753	0.737	0.745	-	-	-
LIBLINEAR	0.741	0.732	0.736	0.553	0.341	0.422
LIBLINEAR†	-	-	-	0.502	0.215	0.301
memorize/closest (baseline)	0.620	0.589	0.604	0.403	0.067	0.115
Phase 2: Includes manual annotations of TIMEX3 and EVENT						
MAX	-	0.843	-	0.823	0.564	0.573
CRF++	0.844	0.843	0.844	-	-	-
LIBLINEAR	0.818	0.818	0.818	0.657	0.418	0.511
LIBLINEAR†	-	-	-	0.562	0.254	0.350
memorize/closest (baseline)	-	0.675	-	0.459	0.154	0.231

Table 4: Relation classification results on the test set. †Indicates official scores before bug correction.

228 were overlapping (but not matching perfectly) with reference annotations, and the remaining 124 errors were due to other reasons. If counting these overlapping errors as true positives instead of false positives, as in a partial match evaluation, significant accuracy improvements could be observed (P: 0.929, R: 0.833, F1: 0.878)[2]. Contributions from last year's TempEval task have pointed out the issue of TIMEX3 annotations inconsistency in the reference standard (Tissot et al., 2015). After examining the 228 overlapping false positive errors further, we noticed through empirical analysis that many were due to either missing or added prepositions (e.g., 'at', 'in', 'for', 'about') and determiners ('a', 'the'). Further examination revealed that, as pointed out by the previous authors, there is an inconsistent trend in the reference standard annotations. For example, the reference standard contains the following TIMEX3 phrases (underlined words indicate words not annotated in the reference standard): "in the past", "in the last three days", "for many years", "for two years", "at this time", "at this time", "about 27 years ago" and "about 30 years ago". These irregularities will make it difficult for any machine learning model to generalize well beyond the given dataset and most likely will indicate overfitting for higher performance models (Velupillai et al., 2015b). The reported inter-annotator agreement for TIMEX3 span annotations of F1 77.4% (Bethard et al., 2015) further supports these assumptions. Therefore, future work should focus on creative ways to deal with this inconsistency and enable more generalizable solutions. Apart from the overlapping errors due to reference standard inconsistencies; other types of errors may indicate room for future improvement. We believe that training multiple classifiers and combining the outputs using ensemble-based approach could yield superior results as manifested from combining predictions of CRF and LIBLINEAR models.

For the DocTimeRel subtask, the CRF-based classification approach also allowed for significant improvements, particularly in phase 2. Table 5 shows the confusion matrix and evaluation scores obtained on the dev set for each category of DocTimeRel relation using CRF++ model when trained on the training set. The final scores achieved (R 83.3%) on the dev set, are comparable to the scores achieved (R 84.3%) on the test set. This allows us to make consistent conclusions about classifier performance on one set (dev) that can be expected to apply on the other set (test). The lowest accuracy (R 48.6%) was observed with the BEFORE/OVERLAP category. A possible explanation for this lower accuracy is the small number of training samples available in this category (2160 instances in the training set out of 38885). The confusion matrix shows that this category gets almost a balanced error rate between the BEFORE (297) and OVERLAP (271) categories. In addition, the highest number of misclassified instances occur in OVERLAP (972) and BEFORE (858) categories where one category is confused for the other. Future work should focus on improving classification in the BEFORE and OVERLAP categories.

The performance achieved using LIBLINEAR models in the CONTAINS relations subtask (F1 42.2%–51.1%) is a significant improvement over last year's attempt using a CRF model (F1 12.3%–

[2] This score was obtained using the `--overlap` option from the official evaluation script.

		SYSTEM				
		AFTER	BEFORE	BEFORE/OVERLAP	OVERLAP	TOTAL
REFERENCE	AFTER	**1686**	157	5	289	2137
	BEFORE	110	**6667**	145	*972*	7894
	BEFORE/OVERLAP	12	297	**548**	271	1128
	OVERLAP	231	858	145	**8579**	9813
	TOTAL	2039	7979	843	10111	20972
	SCORE (P/R/F1)	0.827/0.789/0.807	0.836/0.845/0.840	0.650/0.486/0.556	0.848/0.874/0.861	0.831/0.833/0.831

Table 5: Confusion matrix and scores for each category of DocTimeRel relation obtained on the dev set using CRF++ classifier.

26.0%) (Velupillai et al., 2015c). We think that studying different strategies for candidate pair selection or experimenting with different class weights to reduce effects of negative class predictions could allow for improvement in this subtask. In addition, although we used two separate models to predict relations between event pairs within and between consecutive sentences, we restricted the way we chose candidates across sentences (first and last from current sentence are paired with first and last from next sentence). This restriction was used to avoid an increase in the number of pairs without a relation (i.e., negative class pairs); in addition to the increased computational runtime penalty. However, this means that any candidate pairs spanning across many sentences will be missed by our classifier. This is especially true for some event and time phrases that are usually at the beginning of a sentence (mostly introducing a section header) and act as narrative containers for many events in the next few sentences. For instance, our classifier missed the 'HISTORY' narrative container appearing as part of the section header "PAST MEDICAL HISTORY", which is usually a relation source for many events discussed within the section. One example from the dev set shows that the 'HISTORY' event CONTAINS following events (e.g., medical conditions in a numbered list) spanning from the next first sentence down to the eleventh sentence. Future work could focus on using carefully hand-crafted rules to capture these pairs to increase recall. We think that the most successful approach for this subtask could use hybrid approaches combining rules and machine learning classifiers to improve recall and retain high precision, respectively.

5 Conclusion

Temporal information extraction and reasoning from clinical text remains a challenging task. Our analysis of different machine learning approaches have been informative, and resulted in competitive results for the 2016 Clinical TempEval subtasks. We plan to develop hybrid and ensemble-based approaches in the future to further improve performance on this, and other clinical corpora.

Acknowledgments

We would like to thank the Mayo clinic and the 2016 Clinical TempEval challenge organizers for providing access to the clinical corpus, and arranging NLP shared task.

References

Juan Carlos Augusto. 2005. Temporal reasoning for decision support in medicine. *Artificial intelligence in medicine*, 33(1):1–24, jan.

Steven Bethard, Philip V Ogren, and Lee Becker. 2014. Cleartk 2.0: Design patterns for machine learning in uima. In *LREC*, pages 3289–3293.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6 : Clinical TempEval. pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. 2010. *Temporal information systems in medicine*. Springer Science & Business Media.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning*, 9(2008):1871–1874.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Taku Kudo. 2005. Crf++: Yet another crf toolkit (2005). *Software available at http://crfpp. sourceforge. net*. Accessed: 2010-02-25.

Chen Lin, Dmitriy Dligach, Timothy A. Miller, Steven Bethard, and Guergana K. Savova. 2015. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, page ocv113.

S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–44, jan.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. *URL http://www. chokkan. org/software/crfsuite*. Accessed: 2016-02-25.

James Pustejovsky and Amber Stubbs. 2011. Increasing Informativeness in Temporal Annotation. *Law*, (June):23–24.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2014:218–23.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2009:568–72, jan.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17:507–513.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

William F Styler, Steven Bethard an Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of Computational Linguistics*, 2(April):143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):806–13.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):814–9.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):828–35.

Hegler Tissot, Genevieve Gorrell, Angus Roberts, Leon Derczynski, Marcos Didonet, and Del Fabro. 2015. UFPRSheffield : Contrasting Rule-based and Support Vector Machine Approaches to Time Expression Identification in Clinical TempEval. *Proc. SemEval*, (SemEval):835–839.

Sumithra Velupillai, D Mowery, BR South, M Kvist, and H Dalianis. 2015a. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*, 10(1):183.

Sumithra Velupillai, D. L. Mowery, S. Abdelrahman, L. Christensen, and W. W. Chapman. 2015b. Towards a Generalizable Time Expression Model for Temporal Reasoning in Clinical Notes. In *AMIA 2015 Proceedings*, pages 1252–1259, San Francisco, USA, November. American Medical Informatics Association.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015c. BluLab : Temporal Information Extraction for the 2015 Clinical TempEval Challenge. *Proc. SemEval*, (SemEval):815–819.

Taowei David Wang, Catherine Plaisant, Alexander J Quinn, Roman Stanchak, and Shawn Murphy. 2008. Aligning Temporal Data by Sentinel Events : Discovering Patterns in Electronic Health Records. *CHI 2008 Proceedings Health and Wellness*, pages 457–466.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):849–58.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical data–a review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202, apr.

CHAPTER 4

AUTOMATED EXTRACTION AND CLASSIFICATION OF CANCER STAGE MENTIONS FROM UNSTRUCTURED CLINICAL TEXT FIELDS IN A CENTRAL CANCER REGISTRY

Abdulrahman K. AAlAbdulsalam, MS, Jennifer H. Garvin, MBA, PhD,
Andrew Redd, PhD, Marjorie E. Carter, MS, Carol Sweeny, PhD,
Stephane M. Meystre, MD, PhD

Proceedings of AMIA Informatics Summit, 2018 (submitted).

4.1 Abstract

Cancer stage is one of the most important prognostic parameters in most cancer subtypes. The American Joint Committee on Cancer (AJCC) specifies criteria for staging each cancer type based on tumor characteristics (T), lymph node involvement (N), and tumor metastasis (M) known as TNM staging system. Information related to cancer stage is typically recorded in clinical narrative text notes and other informal means of communication in the Electronic Health Record (EHR). As a result, human chart-abstractors (known as certified tumor registrars) have to search through voluminous amounts of text to extract accurate stage information and resolve discordance between different data sources. This study proposes novel applications of natural language processing and machine learning to automatically extract and classify TNM stage mentions from records at the Utah Cancer Registry. Our results indicate that TNM stages can be extracted and classified automatically with high accuracy (extraction sensitivity: 95.5%–98.4% and classification sensitivity: 83.5%–87%).

4.2 Introduction

Cancer is the second leading cause of death in the United States and recently became the leading cause of death in 21 states, surpassing heart diseases. In the United States, about 595,690 cancer deaths are estimated to have occurred in 2016, which is about 1,630 people per day.[14] The burden of cancer on public health has mobilized national and international institutions to develop strategies to combat, prevent and control cancer.[12,17]

One of the resources vital to the fight against cancer is cancer registries that collect critical information at the population level. Human abstractors, known as Certified Tumor Registrars (CTRs), are tasked with identifying reportable cancer cases and manually collecting data required for cancer registries. This is often a time-consuming and laborious process that is prone to human error and affects quality, completeness and timeliness of cancer registry data. A study based on surveys conducted across European cancer registries as part of the EUROCOURSE project and covering a population of more than 280 million found that the median time to complete case ascertainment for the relevant year was 18 months, with an additional 3-6 months to publish data to national databases .[1,22] Though delay in completion is primarily related to clinical processes to evaluate the patient's extent of disease, reduced time to ascertainment is very desirable. The Maryland Cancer Registry reported 13% of cases with missing staging information .[6] A similar study in the Ottawa Regional Cancer Centre found missing staging information in 10% of lymphoma cases and 38% of breast cancer cases.[20] In a prostate cancer study in Connecticut, about 23% of cases in the registry had incorrectly coded staging information.[7] A study conducted in Los Angeles County Cancer Surveillance Program (CSP) database found that 77% of cases with testicular cancer were coded with inaccurate stage group.[3] While there is a range of accuracy of stage determination, an automated or semiautomated process consisting of a systematic review of relevant text information could potentially improve accuracy.

The American Joint Committee on Cancer (AJCC) manual specifies criteria for staging each cancer site depending on primary tumor characteristics (T), number and location of lymph nodes involvement (N), and metastatic nature (M). Information about the cancer stage is critical for assessing prognosis and selection of treatment plans. Clinical guidelines require clinicians to assign TNM stages prior to initiating any treatment.[2] The *clinical* TNM stage is determined based on the results of physical exams, imaging (such as x-rays or CT scans), and tumor biopsies. The *pathological* TNM stage is determined based on surgery to remove a tumor or explore the extent of the cancer.

Natural Language Processing (NLP) coupled with Machine Learning (ML) are promising technologies to increase the efficiency of cancer registry data abstraction processes. NLP and statistical machine learning have been successfully applied in several medical domains to extract various types of information from clinical text. Their potential for increased efficiency and manual process automation has been demonstrated.[10] In the domain of cancer, several studies showed effectiveness of NLP and ML to mine the electronic health record for cancer-related information from a variety of report types[15] and automatically discover reportable cancer cases based on analysis of pathology records.[4] In the study presented here, we use state-of-the-art Natural Language Processing (NLP) and Machine Learning (ML) to automatically extract TNM stage mentions from patient records collected at the Utah Cancer Registry. The TNM mentions are classified as either pathological or clinical depending on contextual information and will subsequently be used to automatically consolidate stage information and assign a stage group for each cancer case within the registry.

4.3 Related Work

Most previous studies have focused on extracting AJCC TNM stage information exclusively from pathology reports.[5,8,9,11,19] This would not support cancer registry efforts adequately because clinical staging is assigned prior to initiation of treatment and many patients do not immediately undertake resection of tumor and pathology examination. Newer editions of the AJCC TNM manual specifically include separate clinical stage (designated with cT, cN, and cM) and pathological stage (designated with pT, pN, pM) to reflect this time-sensitive staging mechanism. Cancer registries require the use of both clinical and pathological stage information to find the most accurate stage group. The former is primarily based on clinical examination tests and findings (e.g., imaging reports such as CT-scan) and cannot, by definition, be assigned based on information from pathology reports.

McCowan et al.[9] focused their work on extracting T and N stages from lung cancer pathology reports. The pathology reports were first preprocessed to standardize input followed by document-level bag-of-word classifiers to detect relevant reports that contain enough information for the T and N stage classification. They then used a series of rule-based and support vector machine (SVM-based) classifiers at the sentence level to detect phrases with relevant T and N stage information based on factors found in the TNM stage guidelines such as tumor dimension and lymph node involvement. The highest T and N stages detected by the sentence-level classifiers were assigned to each patient. Their approach achieved accuracies of 74.3% and 86.6% for T and N stage classification, respectively, when trained on a dataset of 710 cases, and evaluated on a held-out dataset of 179 cases. The authors used the manually-assigned TNM pathologic stages as the gold standard to measure accuracy of their system. Since the approach to perform stage classification heavily relies on factors associated with lung cancer that were obtained using expert manual annotations, it would be difficult to generalize to other cancer sites without retraining the whole system on new annotations. Nguyen et al.[11] used a similar lung cancer dataset and replaced the machine learning component of McCowan et al. with a rule-based dictionary component. Their approach eliminated the need for expert manual annotations, and achieved comparable accuracies of 73% and 79% for T and N stages, respectively.

Martinez et al.[8] experimented with various machine learning approaches to identify

TNM stages for colorectal cancer in reports obtained from the Royal Melbourne Hospital in Australia. A notable aspect of their experiment was assessment of the generalizability of their methods when using a colorectal cancer dataset from a different institution. Their results showed that accuracy dropped significantly from above 80% down to 50% to slightly above 60% when training and testing on the same corpus versus using corpora from different institutions (cross-corpora) for training and testing. They attribute this drop mainly to differences between the two corpora in expressing TNM labels (e.g., *T1* for staging T). Based on feature selection analysis performed by the authors, these explicit TNM labels are among the top features for good performance within the same corpus, and differences across corpora may lead to inconsistent predictions that could introduce many errors and hamper good performance.

Kim et al.[5] focused on extracting TNM stage information from pathology notes of prostate cancer patients. Using a set of 100 radical prostatectomy specimen reports, they first created a gold standard using two blinded manual reviewers. They then used an NLP system developed to directly match TNM mentions like pT2 and achieved very high accuracies of 99%, 95% and 100% for T, N and M stages, respectively. It is worth mentioning, however, that for the M stage, the dataset was highly skewed with all 100 cases from the randomly selected sample staged as MX.

Warner et al.[19] implemented an NLP system that searched and directly found relevant phrases for summary stage information (i.e., stage I, stage II, early stage, etc.) from the entire EHR available at their institution. They successfully achieved high accuracy (Cohen's kappa of 0.906) when comparing with stages manually determined at the cancer registry, and using a set of 2,323 cancer cases with about 751,880 documents.

Based on the prior scientific work cited above, we used a hybrid approach combining pattern-matching for extraction and supervised machine learning for classification of TNM stage mentions as either pathological or clinical. To the best of our knowledge, our study is the first to report about the extraction of TNM staging information from unstructured text found in records collected at a central Cancer Registry (UCR).

4.4 Methods

4.4.1 Utah Cancer Registry Data

This study is based on data collected at the UCR, which instructs and oversees more than 70 cancer facilities in the state of Utah. Each cancer treatment facility sends records abstracts containing the required data elements for a given cancer patient electronically to the Utah Cancer Registry for each newly diagnosed cancer case. Two types of reports were used for this study. The first type is the North American Association of Central Cancer Registries (NAACCR) abstract record,[16,18] which contains coded information required for reporting to national cancer databases such as the patient age, date of diagnosis, cancer tumor histology, and grade. NAACCR abstracts also contain unstructured text fields that include information such as the patient clinical history, clinical exam results, imaging study descriptions, and any potential staging information. The other type of record is the electronic surgical pathology report also known as E-path. Its content consists of mostly unstructured text fields about the tumor gross pathology, histology, and final diagnoses. Since a patient can be seen in multiple different facilities within a state or have multiple visits within the same facility, there are usually multiple NAACCR and E-path records available for each patient at the Utah Cancer Registry. We refer to these reports as *unconsolidated* records in this study. The role of registrars at the Utah Cancer Registry is to consolidate all information received by the registry for a given cancer case and to produce one final *consolidated* abstract that captures the most accurate information for final reporting to national authorities.

4.4.2 Reference Standard

For development and evaluation of our system, a random subset of 100 cancer cases was selected from three different cancer primary sites (see Table 4.1): Colon, Lung and

Table 4.1: Document types and counts for the corpus used in this study. N = number of patient cases.

Record Type	**QCSET (N=60)**	**ABSTRACTION (N=240)**	**TOTAL (N=300)**
NAACCR	72	286	358
E-path	113	339	452
TOTAL	185	625	810

Prostate cancers with 300 cases in total. These three primary sites are among the most prevalent cancer types at the UCR and could therefore benefit the most from case review and consolidation automation. The text fields from NAACCR (see Table 4.2) and e-path records for these 300 cases constituted the corpus for this study. Note that since each case may have multiple records (3 on average), the corpus contains far more documents than selected cancer cases. In our case, the corpus consisted of 810 NAACCR and e-path records as shown in Table 4.1. All text fields in these records were manually annotated for mentions of TNM staging information. Two human annotators who are certified tumor registrars conducted the annotation independently, and a third domain expert participated in the process for adjudication of differences between annotators when necessary. The annotation task was initiated by going through preliminary practice rounds in which annotators were given the same set of 25 documents to annotate followed by team meetings where agreement was discussed, and annotation guidelines revised to clarify ambiguous examples found during preceding practice sessions. Once an adequate level of agreement ($\kappa = 0.81$) was observed and good understanding of the annotation task was achieved, we started the quality control phase in which a small subset (QCSET) of 20 cases from each cancer

Table 4.2: NAACCR column names and numbers for the free text fields used in the study.

NAACCR Item Number #	Text Field Name
2520	Text–Dx Proc–PE
2530	Text–DX Proc–X-ray/scan
2540	Text–DX Proc–Scopes
2550	Text–DX Proc–Lab Tests
2560	Text–DX Proc–Op
2570	Text–DX Proc–Path
2580	Text–Primary Site Title
2590	Text- Histology Title
2600	Text–Staging
2610	RX Text–Surgery
2620	RX Text–Radiation (Beam)
2630	RX Text–Radiation Other
2640	RX Text–Chemo
2650	RX Text–Hormone
2660	RX Text–BRM
2670	RX Text–Other
2680	RX Text–Remarks
2690	Text–Place of Diagnosis

site (60 cases total) from the reference standard was selected for double-annotation. The interannotator agreement was calculated using Cohen's kappa statistic and results are shown in Table 4.3. Disagreements in the QCSET were mostly due to either a missed TNM value mention by one annotator or discrepancies in the timing attribute. Given the excellent interannotator agreement observed, the remaining documents in our corpus (ABSTRACTION) were each annotated by one annotator only. The annotation project and reference standard development was managed using the WebAnno tool.[21] The annotation schema included three categories of information to be annotated corresponding to T, N and M stage mentions. Each TNM stage mention was annotated with the following attributes:

- **Stage**: The AJCC stage designation (e.g., T1, N1b).

- **Timing**: This attribute is used to indicate if the staging is *clinical* or *pathological* as per the rules of the AJCC manual and according to the context of the mention.

- **Negation**: The value to indicate if a TNM mention is within the scope of a negated context. The possible values are: *affirmed* (default, or most mentions), *negated*, and *possible*. Most mentions will have an affirmed value. The *possible* was selected when there was hedging involved within the context of mention.

- **Temporality**: This attribute is used to capture historical or future mentions that do not necessarily represent current mentions valid at the point in time when the mention was stated at the patient record. The three possible values are: *current* (default, or most mentions), *historical*, and *hypothetical* (future mentions).

- **Subject**: This is used to capture TNM mentions that are related to family relatives or others who are not the patient himself. Possible values include: *patient* (default), and *other*.

Table 4.3: Interannotator agreement by document type in the QCSET. Method is Cohen's Kappa for 2 raters.

Document Type	**Mentions annotated by both raters**	**Kappa**	**p-value**
e-path	60	0.658	< 0.001
NAACCR abstract	125	0.9009	< 0.001
All	185	0.8129	< 0.001

After completion of the reference standard development, we found that almost all TNM mentions were affirmed, recent and related to the patient (only 1 was negated, 3 were historical and 2 were related to someone other than the patient). Therefore, we decided to exclude negation, temporality and subject attributes from further training and analysis. Table 4.4 presents TNM annotations added by annotators in the reference standard. The reference standard was divided randomly into 3 subsets: Train (50%), Development (17%) and Test (33%).

4.4.3 NLP and ML Systems

Figure 4.1 shows the complete NLP and ML system developed for this study. The preprocessing components were adapted from cTAKES[13], a general clinical NLP application. Each text field is broken into smaller units (sentences and tokens such as words, numbers, and punctuation) and assigned parts-of-speech tags by preprocessing modules for further analysis. Both rule-based (pattern-matching) and ML-based (Conditional Random Fields) approaches were utilized for TNM mentions extraction and classification. The pattern matching component was developed using regular expressions based on human annotations from the training data to achieve high sensitivity. The Conditional Random Fields (CRF) component was developed using the CRFsuite package within ClearTK machine learning libraries. The Java-based Apache UIMA framework was used as the main development framework for this project.

The development of NLP included the following system variants:

1. **REGEX**: based on direct pattern matching using regular expressions for TNM mentions, and rules for classifying timing attribute, as follows,

   (a) If TNM mention has prefix letter 'c' then clinical else if it has prefix letter 'p' then pathological.

Table 4.4: TNM mentions extracted by annotators from corpus used as reference standard.

Data Subset	Records	T	N	M	Total TNM annotations
Train (50%)	405	235	192	86	513
Development (17%)	135	85	73	27	185
Test (33%)	270	139	119	52	310
All	810	459	384	165	1008

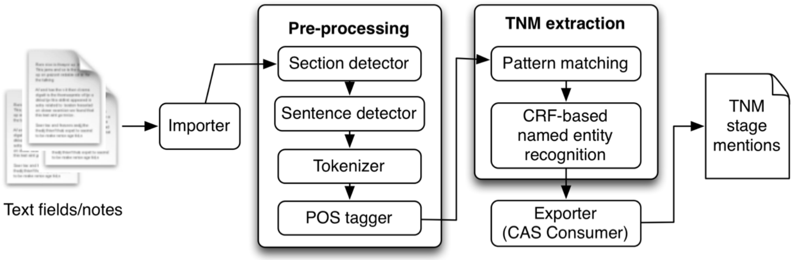


Figure 4.1: NLP and ML application high-level architecture.

(b) If TNM mention has no 'c' or 'p' prefixes and TNM mention was extracted from pathology record then pathological, otherwise if extracted from NAACCR abstract then clinical.

2. **CRF**: based on a Conditional Random Fields (CRF) machine learning algorithm.

3. **REGEX-CRF**: hybrid system that combined regular expression output with CRF algorithm classification of timing attribute.

4.5 Results

Tables 4.5 and 4.6 detail the results of our evaluation of the NLP system variants. Reported metrics include: precision (equivalent to positive predictive value), recall (sensitivity) and the F1-measure (harmonic mean of precision and recall). We report evaluations with both subsets: development and test. Results with the development subset were obtained by training with the train subset only while results with the test subset were obtained after training on both the train and development subsets. We used two evaluation approaches to compare the NLP system output with our reference standard: strict and

Table 4.5: TNM mentions extraction results.

Evaluation Method	System	Development Set			Test Set		
		Precision	Recall	F1-measure	Precision	Recall	F1-measure
Strict match	REGEX	0.926	0.946	0.936	0.890	0.884	0.887
	CRF	0.952	0.859	0.903	0.923	0.845	0.882
Partial match	REGEX	0.958	**0.984**	**0.971**	0.961	**0.955**	**0.958**
	CRF	**0.988**	0.897	0.940	**0.989**	0.906	0.946

Table 4.6: Pathological and clinical TNM classification results. P – Precision, R – Recall.

Evaluation Set	System	Pathological			Clinical			Overall		
		P	R	F1	P	R	F1	P	R	F1
Development	REGEX	0.952	0.688	0.798	1.000	0.025	0.049	0.529	0.543	0.536
	REGEX-CRF	0.901	0.889	0.895	0.681	0.800	0.736	0.847	**0.870**	0.858
Test	REGEX	0.934	0.536	0.681	1.000	0.051	0.097	0.386	0.384	0.385
	REGEX-CRF	0.859	0.896	0.877	0.793	0.704	0.746	0.841	**0.835**	0.838

partial matches. The former requires exact matching of the TNM mention sequence of characters predicted by the NLP system with the corresponding mention in the reference standard. The latter relaxes this restriction by allowing overlapping TNM mentions to be counted as true positives. For instance, if the NLP system extracted the mention “T1” and the reference standard had “cT1” then this would count as a true positive partial match. Table 4.5 shows the results of TNM mentions extraction. Table 4.6 shows the results of classifying the TNM mentions to pathological and clinical. Note the REGEX-CRF system uses REGEX to extract TNM mentions and CRF to classify them to pathological and clinical mentions.

Both the REGEX and CRF versions of the NLP system achieve comparable performance when detecting mentions of TNM staging (F1: 94.0%–97.1%), but the CRF version reached much higher accuracy when predicting the timing (clinical or pathological) attribute (F1: 83.8%–85.8%). In general, the REGEX (rule-based) version retrieved slightly more correct TNM mentions (i.e., had higher recall: 88.4%–98.4%) while the CRF version retained a higher precision (Precision: 92.3%–98.9%). The CRF version also performed better than the REGEX version when classifying the timing attribute. Our analysis of partial matching results indicate that the NLP system tends to miss words like “clinical” and “PATH” preceding TNM mentions as included by our annotators in the reference standard. This is despite the fact that our annotation guidelines did not in particular include specific instructions to the annotators on selecting contextual words preceding TNM mentions as in these cases. The NLP system was able to partially capture these longer multiword TNM mentions found in the reference standard. The REGEX NLP system achieved very high sensitivity with partial matching, indicating that this approach is practical for extracting TNM mentions from NAACCR and E-path records. The hybrid REGEX-CRF NLP system combined the output of the REGEX version with the classification of timing by the CRF module, and achieved the highest F1-measure overall (F1: 83.8%–85.8%).

4.6 Discussion

Our effort outlined above showed that extraction of TNM mentions from unstructured text fields within records collected at the central Utah Cancer Registry can be automated with reasonable accuracy. One source of errors that was eliminated earlier during our development was matching of ‘TX’ mentions by the REGEX NLP system. Despite slight increase in sensitivity, there were numerous false positives and a decrease in precision when including this pattern since it could be confused with the commonly used ‘TX’ abbreviation for ’treatment’ in this corpus (see Table 4.7, for examples). For this reason, we decided to exclude the pattern for ‘TX,’ especially since patient cases with no T stage mentions extracted from their records can be assigned ‘TX’ stage by default. Most strict matching errors were caused by missing contextual words preceding TNM mentions such as “CLINICAL” and “PATHOLOGIC”. When considering partial matches, many errors were caused by spurious matching within alphanumerical terms such as matching ‘T0’ or ‘N1’ in “**T0**012-9071” and “**N1**3-129”. Other errors were due to confusion of text mentions related to MRI scans such as ‘T2’ inside the statement: “SUBTLE AREA OF FOCAL **T2** SIGNAL LOSS”. Similarly, incorrectly matching ‘T1’ within a biomarker phrase as in the sentence:“weakly positive for W**T1**”. A third source of errors was the use of capital letter ‘O’ instead of the digit ‘0’ in some TNM mentions such as “NO” and “MO” instead of “N0” and “M0” stages, respectively. These errors could be addressed by including more contextual, lexical and character shape features to enable disambiguation and improve sensitivity while maintaining high precision.

Although regular expressions were more robust for extracting TNM mentions, our range of features used with the CRF classifier were still limited and potential improvements may be observed if other more sophisticated feature patterns were used such as character N-grams.

Table 4.7: Example statements containing the ‘TX’ abbreviation.

Statements with TX abbreviations
DISCUSSED PALLIATIVE TX W/ CARBO/TAXL ...
NEW LUNG CANCER F/U & TX ...

In addition, other machine learning algorithms could yield better performance than CRF and further investigation is required.

Results reported here were validated with patient records from various healthcare organizations (local and regional hospitals) in the state of Utah. We believe that despite potential differences in documentation style and use of linguistic patterns, the proposed NLP and ML systems were able to extract TNM mentions with high accuracy comparable to manual abstraction by humans. To investigate questions about the distribution of TNM mentions extracted across sites, the number of TNM mentions found for each patient case, and the percentage of patients without any TNM mentions in their records, we applied the REGEX NLP system to a selected set of 11,180 NAACCR and e-path records for a population of 4,117 patient cases available from the UCR database. Table 4.8 outlines the number of TNM mentions extracted across cancer sites. There were 14,560 mentions extracted in total from all case records. In general, more patients had no M stage mentioned in their records, followed by N stage mentions, and finally T stage mentions. Across the three primary cancer sites, colon cases tended to have more TNM mentions in their records than prostate or lung cancer cases.

The number of TNM mentions extracted from patient records was distributed as shown in Figure 4.2. On average, there were about 5 mentions extracted per patient. The distribution is right skewed with a majority of patients having less than 10 mentions and then gradually fewer patients having more than 10. At the extreme right were patients with more than 30 TNM mentions. Note that this average excludes patients who have no TNM

Table 4.8: Count of TNM mentions extracted from a selected set of patient cases.

Site	**TNM**	**count**
Colon	T	2814
	N	2635
	M	1012
Lung	T	1615
	N	1407
	M	634
Prostate	T	2341
	N	1409
	M	693
Total		14560

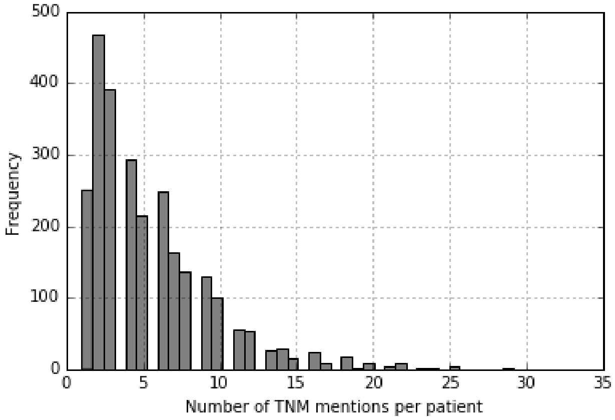


Figure 4.2: Frequency of TNM stage mentions extracted per patient.

stage mentions in their records. Only patients with TNM stage mentions extracted from their records were considered.

There were about 6,485 records (out of a total of 11,180) with no TNM mentions, belonging to 1,443 patient cases (out of 4,117). When considering each T, N and M mention individually, about 37% of the patients had no T stage mentions (1558/4117), 44% had no N stage mentions and 63.5% had no M stage mentions.

4.7 Conclusion

The study presented here showed that automated extraction of TNM stage information using NLP and ML approaches could achieve high accuracy, at levels comparable with manual abstraction by humans. In a future study, we plan to use the NLP pipeline developed for TNM stage information extraction to then perform cancer stage consolidation at the patient level for cases from the three primary cancer sites included in this study. The automated stage consolidation will be compared with the consolidated stages assigned by human registrars manually in the central registry. Our aim is to eventually assess whether NLP and machine learning could be implemented with sufficient accuracy to automatically consolidate cancer stage and support the work of cancer registrars.

4.8 Acknowledgement

The authors would like to thank SuAnn McFadden, Kay McCandless and Jacque Clarken at the Utah Cancer Registry for spending the time to participate in the annotation effort and creation of the reference standard. This work has been supported by contract number HHSN261201300017I from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) program.

4.9 References

[1] J. W. Coebergh, C. van den Hurk, S. Rosso, H. Comber, H. Storm, R. Zanetti, L. Sacchetto, M. Janssen-Heijnen, M. Thong, S. Siesling, et al., *Eurocourse lessons learned from and for population-based cancer registries in europe and their programme owners: Improving performance by research programming for public health and clinical evaluation*, European Journal of Cancer, 51 (2015), pp. 997–1017.

[2] S. B. Edge and C. C. Compton, *The American Joint Committee on Cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM*, Annals of Surgical Oncology, 17 (2010), pp. 1471–1474.

[3] K. D. Faber, V. K. Cortessis, and S. Daneshmand, *Validation of surveillance, epidemiology, and end results tnm staging for testicular germ cell tumor*, in Urologic Oncology: Seminars and Original Investigations, vol. 32, Elsevier, 2014, pp. 1341–1346.

[4] D. A. Hanauer, G. Miela, A. M. Chinnaiyan, A. E. Chang, and D. W. Blayney, *The registry case finding engine: An automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes*, Journal of the American College of Surgeons, 205 (2007), pp. 690–697.

[5] B. J. Kim, M. Merchant, C. Zheng, A. A. Thomas, R. Contreras, S. J. Jacobsen, and G. W. Chien, *Second prize: A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports*, Journal of Endourology, 28 (2014), pp. 1474–1478.

[6] A. C. Klassen, F. Curriero, M. Kulldorff, A. J. Alberg, E. A. Platz, and S. T. Neloms, *Missing stage and grade in maryland prostate cancer surveillance data, 1992–1997*, American Journal of Preventive Medicine, 30 (2006), pp. S77–S87.

[7] W.-L. Liu, S. Kasl, J. T. Flannery, A. Lindo, and R. Dubrow, *The accuracy of prostate cancer staging in a population-based tumor registry and its impact on the black-white stage difference*, Cancer Causes and Control, 6 (1995), pp. 425–430.

[8] D. Martinez, L. Cavedon, and G. Pitson, *Stability of text mining techniques for identifying cancer staging*, in Louhi, The 4th International Workshop on Health Document Text Mining and Information Analysis, NICTA, Canberra, Australia, 2013.

[9] I. A. McCowan, D. C. Moore, A. N. Nguyen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and M.-J. Fry, *Collection of cancer stage data by classifying free-text medical reports*, Journal of the American Medical Informatics Association, 14 (2007), pp. 736–745.

[10] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, et al., *Extracting information from textual documents in the electronic health record: A review of recent research*, Yearb Med Inform, 35 (2008), p. 44.

[11] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist, *Symbolic rule-based classification of lung cancer stages from free-text pathology reports*, Journal of the American Medical Informatics Association, 17 (2010), pp. 440–445.

[12] D. M. Parkin, *The evolution of the population-based cancer registry*, Nature Reviews Cancer, 6 (2006), p. 603.

[13] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, *Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications*, Journal of the American Medical Informatics Association, 17 (2010), pp. 507–513.

[14] R. L. Siegel, K. D. Miller, and A. Jemal, *Cancer statistics, 2016*, CA: Cancer Journal for Clinicians, 66 (2016), pp. 7–30.

[15] I. Spasić, J. Livsey, J. A. Keane, and G. Nenadić, *Text mining of cancer-related information: Review of current status and future directions*, International Journal of Medical Informatics, 83 (2014), pp. 605–623.

[16] A. Stewart, A. Hurlbut, L. Havener, F. Michaud, S. Capron, L. Ries, et al., *North American Association of Central Cancer Registries NAACCR 2006 implementation guidelines and recommendations.* `http://www.naaccr.org/`, 2012.

[17] B. Stewart, C. P. Wild, et al., *World Cancer Report 2014*, Health, (2017).

[18] M. Thornton and L. OConnor, *Standards for cancer registries volume II: Data standards and data dictionary, record layout version 12.2*, North American Association of Central Cancer Registries, (2012).

[19] J. L. Warner, M. A. Levy, M. N. Neuss, J. L. Warner, M. A. Levy, and M. N. Neuss, *Recap: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data*, Journal of Oncology Practice, 12 (2015), pp. 157–158.

[20] J. C. Yau, A. Chan, T. Eapen, K. Oirourke, and L. Eapen, *Accuracy of the oncology patients information system in a regional cancer centre*, Oncology Reports, 9 (2002), pp. 167–169.

[21] S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann, *Webanno: A flexible, web-based and visually supported system for distributed annotations.*, in ACL (Conference System Demonstrations), 2013, pp. 1–6.

[22] R. Zanetti, I. Schmidtmann, L. Sacchetto, F. Binder-Foucard, A. Bordoni, D. Coza, S. Ferretti, J. Galceran, A. Gavin, N. Larranaga, et al., *Completeness and timeliness: Cancer registries could/should improve their performance*, European Journal of Cancer, 51 (2015), pp. 1091–1098.

CHAPTER 5

MACHINE LEARNING TO AUTOMATE CANCER STAGE CONSOLIDATION IN A CENTRAL CANCER REGISTRY

Abdulrahman K. AAlAbdulsalam, MS, Jennifer H. Garvin, MBA, PhD,
Andrew Redd, PhD, Marjorie E. Carter, MS, Carol Sweeny, PhD,
Stephane M. Meystre, MD, PhD

5.1 Abstract

Consolidating cancer stage from multiple records is one of the primary tasks performed by central cancer registries. A team of certified tumor registrars (CTR) conducts the consolidation by manually reviewing records received from multiple sources for each newly diagnosed cancer case. The large volume of cases handled by central registries and the complexity of staging guidelines make staging one of the barriers to reducing the time delay between diagnosis and reporting for national surveillance data.

The objective of this study is to implement and evaluate Natural Language Processing (NLP) and Machine Learning algorithms to automate cancer stage consolidation.

Records collected at the Utah Cancer Registry (UCR) for patients with colon, lung, or prostate cancers were used for this study. UCR receives multiple records for each cancer case containing clinical and pathological information as structured data and text fields. We annotated the source documents, then developed an NLP application to find and extract mentions of the three components of stage, i.e., tumor size (T), node involvement (N), and metastasis (M). Multiple machine learning classifiers were applied to consolidate stage. The consolidated T, N, and M stages were used to derive a stage group for each case. Results were compared to stages assigned manually by experienced registrars as a gold standard.

Consolidation of the M stage for the three cancer sites reached very high accuracy (93.9%96.8%) compared to the manually reviewed gold standard, whereas consolidation of T and N stages varied for different cancer sites. Best accuracy for T and N was observed for colon cancer cases (83.6%91.2%), followed by prostate cancer cases (73.5%81.4%) and lowest for lung cancer cases (60.4%71.1%). Deriving a stage group from consolidated TNM stages reached high accuracy for colon cancer (88.4%) followed by lung cancer (84.5%) while accuracy for prostate cancer was moderate (67.1%).

Automatic consolidation of cancer stage using NLP and machine learning can reach high accuracy for some cancer sites and may be practical and useful in the context of manual human review assistance. Future studies may focus on additional feature engineering and extraction of medical concepts to improve consolidation accuracy.

5.2 Background and Significance

Population-based cancer registries play a critical role in the fight against cancer, collecting cancer information, disseminating vital statistics, providing data for epidemiological studies, and facilitating planning, control and prevention of the disease.[10] For example, the Surveillance, Epidemiology, and End Results (SEER) program at the U.S. National Cancer Institute (NCI) publishes annual statistics about incidence, prevalence and survival patterns of cancer in the United States.[11] Cancer surveillance data compiled by state-wide or regional central cancer registries are the basis for national cancer surveillance. Central registries depend on trained certified tumor registrars (CTRs) to manually curate electronic records of newly diagnosed cancer cases received from a variety of sources and to perform coding of specified required data elements. Increased automation of this process and reduced reliance on manual coding could help address concerns about timeliness of cancer registry reporting and shortage of qualified CTRs.

Cancer staging information is a crucial component of surveillance data for assessing trends and for research. The current staging schemes specified by the American Joint Commission on Cancer (AJCC) and International Union Against Cancer (UICC) include tumor size (T), node involvment (N), metastatis (M), and overall stage group for each case. Assignment of AJCC or UICC stage involves complex coding schemes for each cancer site.

Stage may be missing or incorrectly assigned in records submitted to a central cancer registry. The Maryland Cancer Registry reported 13% of cases with missing staging information because “data are either not clinically ascertained or not successfully abstracted”.[6] Cancer stage information may remain inadequate after central registry adjudication; within the SEER program, the proportion of cases that are unstaged has declined over recent decades but remains high for certain cancer sites.[4] A study of 60 central cancer registries in Europe found that only about a third of them provided stage at diagnosis of good quality, with cases with missing TNM staging information ranging from 0% (breast cancer cases in the Naples, Italy cancer registry) to 95% (lung cancer in the Wales, UK cancer registry).[9] In the New Zealand Cancer Registry, the proportion of invasive breast cancer cases with unknown or inaccurate staging information reached 12%, when all but one of these cases had staging information in the source cancer record.[13] In a prostate cancer study in Connecticut, about 23% of cases in the registry had incorrectly coded staging information.[7] A study conducted in Los Angeles County Cancer Surveillance Program (CSP) database found that 77% of cases with testicular cancer were coded with inaccurate stage group.[3]

Central cancer registries in the U.S. now receive a large proportion of cancer data in electronic form, creating the opportunity for automated abstraction of cancer data .[8,14] In a previous study, we developed and applied NLP and machine learning methods and resources to extract TNM staging information from clinical text found in records at the Utah Cancer Registry (UCR).[1] The present study focuses on applying machine learning algorithms to automatically consolidate extracted TNM stage mentions combined with other structured data into one final stage. The general objective of this study is to assess the extent to which machine learning methods can be used to perform consolidation of cancer case data, and provide insight to future large scale efforts to implement information technology strategies that can address the challenges of increasing costs and sustained data quality. More specifically we aim to:

1. Develop algorithms for consolidating cancer TNM stage information using machine learning.

2. Validate the performance of machine learning algorithms for TNM stage consolidation through comparison to the consolidation decisions made by cancer registrars.

5.3 Materials and Methods

5.3.1 Central Cancer Registry Data

This study is based on data collected at the Utah Cancer Registry (UCR), the central cancer registry for the state of Utah and a SEER program registry. UCR receives reports from more than 70 facilities that treat cancer in the state of Utah. Two types of reports were used for this study. The first type is the North American Association of Central Cancer Registries (NAACCR) abstract record[15,17] which contains information that is compiled by CTRs employed by treating hospitals. Structured fields include the patient age, date of diagnosis, cancer tumor histology, grade, and stage. NAACCR abstracts also contain unstructured text fields that include information such as the patient clinical history, clinical examination results, and imaging study descriptions. The other type of record is the electronic surgical pathology report also known as 'e-path'. Its content includes unstructured text fields about tumor gross pathology, histology, and final diagnoses. A patient may have one or more surgical procedures for diagnosis and treatment of cancer that result in multiple ”e-path” reports to the registry, and the patient may be seen in one or more treatment facilities that each report a NAACCR abstract to the central registry. We refer to these reports as **unconsolidated** records in this study. Registrars at the central registry consolidate and review information received for a given cancer case and produce one final **consolidated** abstract that captures the most accurate information. The final consolidated stage was used as the reference standard for the machine learning study.

5.3.2 Study Population

The study included cases diagnosed between January 1, 2011 and December 31, 2014 with colon, prostate, or lung primary cancer sites. We focused on cases with invasive behavior and on adults aged 20 or more at diagnosis. There were a total of 8,189 cases meeting these eligibility criteria. Of these, 622 were excluded because of missing electronic records (neither e-path nor NAACCR abstract). An additional 12 cases were excluded for histologies that could not be staged. This study was approved by the University of Utah IRB.

A large proportion of the eligible prostate cancer cases were diagnosed with T1/T2, N0, and M0 (Table 5.1). We included a random sample of 500 T1/N0/M0 and 500

Table 5.1: Invasive cancer cases diagnosed in Utah, aged 20 or older at diagnosis. Diagnosed 2011-2014, by cancer site and AJCC derived TNM stage.

	Eligible Cases				Selected Cases			
	Colon	Lung	Prostate	Total	Colon	Lung	Prostate	Total
T0	58	13	6	77	58	13	6	77
T1	308	393	1354	2055	308	393	646	1347
T2	199	501	2418	3118	199	501	869	1569
T3	679	343	486	1508	679	343	486	1508
T4	376	389	52	817	376	389	52	817
TX	149	322	143	614	149	322	143	614
N0	984	746	3741	5471	984	746	1484	3214
N1	392	178	140	710	392	178	140	710
N2	279	605	0	884	279	605	0	884
N3	0	240	0	240	0	240	0	240
NX	114	192	578	884	114	192	578	884
M0	1416	974	4254	6644	1416	974	1997	4387
M1	353	987	205	1545	353	987	205	1545
Total	1769	1961	4459	8189	1769	1961	2202	5932

T2/N0/M0 prostate cases to reduce class imbalance. We included all cases with other TNM combinations for prostate cancer, and we selected all eligible cases for colon and lung cancers. This resulted in 5,932 cases selected for the machine learning task. Table 5.2 shows the distribution of these cases across the three cancer sites considered for this study, along with the demographic distribution of the selected cases by cancer site.

Upon further review of these cases, we identified several patients who had more than one cancer diagnosis among the lung and colon cases. We then restricted the sampled cases to include one cancer site per person. Our final analysis set includes 5915 cases (1760 colon cancer, 1953 lung cancer, 2202 prostate cancer). These cases were divided into 3 subsets for the machine learning task as shown in Table 5S.1. The total number of e-path and NAACCR records available for the 5,915 cancer cases selected for the study were 8,850 and 7,168, respectively, reaching 16,018 total records available for this study.

Table 5.2: Selected cases for inclusion in machine learning project, SEER reportable cases diagnosed 2011-2014 with colon, lung, or prostate cancers.

	Colon	**Lung**	**Prostate**	**Total**
Selected cases	1769	1961	2202	5932
Year of Diagnosis				
2011	434	520	583	1537
2012	421	529	522	1472
2013	456	411	561	1428
2014	458	501	536	1495
Age				
20 to 64	756	674	906	2336
65 to 74	458	679	864	2001
74 or older	555	608	432	1595
Race				
White	1675	1871	2116	5662
Black	26	21	23	70
American Indian or Alaskan Native	17	7	5	29
Asian or Pacific Islander	50	61	48	159
Other or Unknown	1	1	10	12
Ethnicity				
Hispanic	117	126	94	337
Non-Hispanic	1647	1833	2070	5550
Unknown	5	2	38	45
Sex				
Male	904	1003	2202	4109
Female	865	958	0	1823

5.3.3 Consolidated TNM Stages as Gold Labels

Cancer cases selected for the study were assigned an unconsolidated TNM stage by local reporting facilities and then were rectified and assigned a final consolidated stage based on most accurate information by registrars at the Utah Cancer Registry. The consolidation of stage for each patient is an internal process to the Utah Cancer Registry and is usually based on the multiple records received from different sources. Since the registrars at the Utah Central Cancer Registry have access to every record from each facility visited by the patient, they are able to assign the most accurate stage based on available complete information and patient history. Therefore, we will consider the final consolidated stage as the gold standard for this study.

5.3.4 TNM Stage Extraction

A random sample of 300 cases (100 per cancer site included in the study) was selected for annotation and reference standard development. Experienced CTRs annotated the text fields for mentions of T, N and M stages and other relevant contextual information. An NLP application was developed and evaluated based on this reference standard to automatically extract TNM mentions. The methods and results of this system are reported in a different paper.[1] Final evaluation showed that a rule-based NLP system achieved very high accuracy for the extraction of TNM stage mentions (sensitivity: 95.5%–98.4%). The NLP application was applied to the records of all cases used in the current study and the extracted TNM stage mentions from patient records were used as input features to the machine learning algorithms developed for automatic consolidation in this study.

5.3.5 NLP and Machine Learning Application

Figure 5.1 highlights the NLP and machine learning application developed for the study. The application consists of two main components: TNM extraction and TNM consolidation. The TNM extraction component was developed and evaluated based on reference standard of 300 cases (100 per site), and more details and results are reported in a separate study.[1] The preprocessing components were adapted from cTAKES[12] clinical processing pipeline to produce lexical and syntactic features for the TNM extraction. The consolidation of various information from the unconsolidated NAACCR and e-path records into one final TNM stage was implemented using Support Vector Machine (SVM) available from scikit-learn python

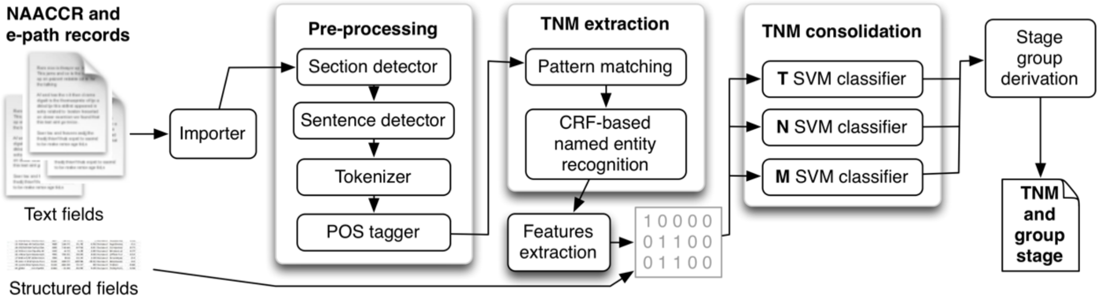


Figure 5.1: NLP and machine learning application high-level architecture.

library (LinearSVC). The feature extraction component produces necessary sparse matrix representation for the SVM learning task. Both structured data (e.g., age, histology, grade) and TNM mentions extracted from unstructured text fields of each patient records were used as input features for the learning task, with the consolidated TNM stage (manually verified by the registrars at the Utah Cancer Registry) as the target label for classification.

5.3.6 Baseline System

A baseline system was developed to consolidate TNM stage mentions extracted by the NLP system. The baseline uses two simple rules:

1. Given multiple T, N and M mentions extracted from multiple records belonging to the same patient, assign the highest numerical stage (selecting the highest value for TNM is the same rule of thumb a certified tumor registrar would use for case consolidation).

2. If the patient has no respective TNM stage mentions, then use the most frequent stage for each T, N, and M as supported by stage counts from training data, for instance, use stage N0 when assigning N stage since it is the most prevalent stage for N in the training data (see TNM stage counts from Table 5.1).

5.3.7 Features from NAACCR Records

Coded and structured data fields from the unconsolidated NAACCR records for each patient were used to train the SVM classifiers used for final stage consolidation. Each data field is a numerical code assigned by registrars after reviewing the patient data. NAACCR standard and data dictionary defines the specification and scope of each data element captured in the NAACCR records. Since each patient may have multiple unconsolidated

NAACCR records, we implemented a procedure that selects the highest coded value for each data field used as feature when multiple numerical codes exist. If any data field is empty for a patient we use -1 to indicate a missing code for this data.

5.3.8 Algorithms and Features Optimization

Large margin classification based on SVM has demonstrated superior results in many studies and recent NLP challenges.[2,16] However, due to the novelty of the current task, we experimented with other well-known classification algorithms in the literature. The scikit-learn library was used to systematically compare performance of various algorithms such as Decision Trees, Naive Bayes, K-Nearest-Neighbors, Random Forests and ensemble-based voting classifier. In addition, we performed a systematic feature selection and parameter tuning using grid search to find the most effective values for the SVM parameters. Feature selection was guided by a stepwise evaluation with progressive addition of features to finally obtain the best feature subset for each cancer site and TNM combination.

5.3.9 Stage Group Derivation

The AJCC rules and guidelines provide stage group classification (e.g., I, II, III and IV) definitions based on TNM stage information. We used these (as found in the AJCC 7th edition manual) for each primary cancer site to derive the stage group based on the predicted TNM labels automatically assigned by our NLP and ML application. We then compared the derived stage group for each case to the stage group assigned by UCR registrars to assess accuracy. An example of stage group classification based on the AJCC manual for the three cancer sites considered in this study is shown in Table 5S.2.

5.3.10 Evaluation Methods

Multiple SVM classifiers were developed to consolidate each of T, N and M staging labels for each of the primary cancer sites separately (9 total SVM classifiers). All initial experiments were conducted on the training and development subsets using cross validation approach. The most effective algorithm, features and parameter settings were selected for final evaluation on the test subset that was heldout from the system developer and only released during the final phase of the study. The predictions on the test set were submitted to an independent statistician who participated in the study to perform final assessment.

The 7th edition of AJCC TNM manual adds a layer of complexity to the TNM classification system by introducing a more refined staging hierarchy with alphabetical subdesignation to the existing numerical stages such as T1a and N2b. We grouped TNM labels assigned by registrars into a more coarse-grained numerical level (essentially converting the more refined alphabetical stages such as T1a and T1b into T1) and then trained classifiers to predict numerical stages only. Similarly, the stage group labels in the consolidated records for each case were converted to numerical level stages (e.g., IIA and IIB into II) to be used for final evaluation.

Since each patient must be assigned a final consolidated TNM and stage group labels, we used accuracy (number of correct classifications divided by number of total cases) as the primary metric to measure performance during development. We report the mean and standard deviations of accuracy obtained from 3-fold cross validation on the training and development subsets. We report final assessments on the test subset using accuracy (percent agreements), and Cohen's kappa (κ) metrics.

5.4 Results

5.4.1 Consolidation with Baseline System

Table 5S.3 shows the accuracy obtained (mean and standard deviations) from applying the baseline system to the 70% training data using 3-fold cross validation approach. This simple baseline achieves accuracy higher than 80% when consolidating N and M for colon, and M for prostate. Consolidation of T for colon, and N and M for prostate achieves lower accuracy level below 70%. Consolidation of T, N, and M stages for lung cancer achieves low accuracy below 60%.

5.4.2 Consolidation with Various Machine Learning Algorithms

To compare various machine learning algorithms and apply the most accurate of them for automatic TNM consolidation, we used the python scikit-learn library. The plots in Figure 5.2 show accuracy levels obtained by running each machine learning algorithm on the training subset using 3-fold cross validation. Each subfigure contains the *mean* accuracy performance from 3-fold on the y-axis and the classification algorithm used on the x-axis. The plots in each subfigure show accuracy for all cancer sites, lung, colon and prostate, from top to bottom, respectively. The left column shows accuracy when using extracted TNM

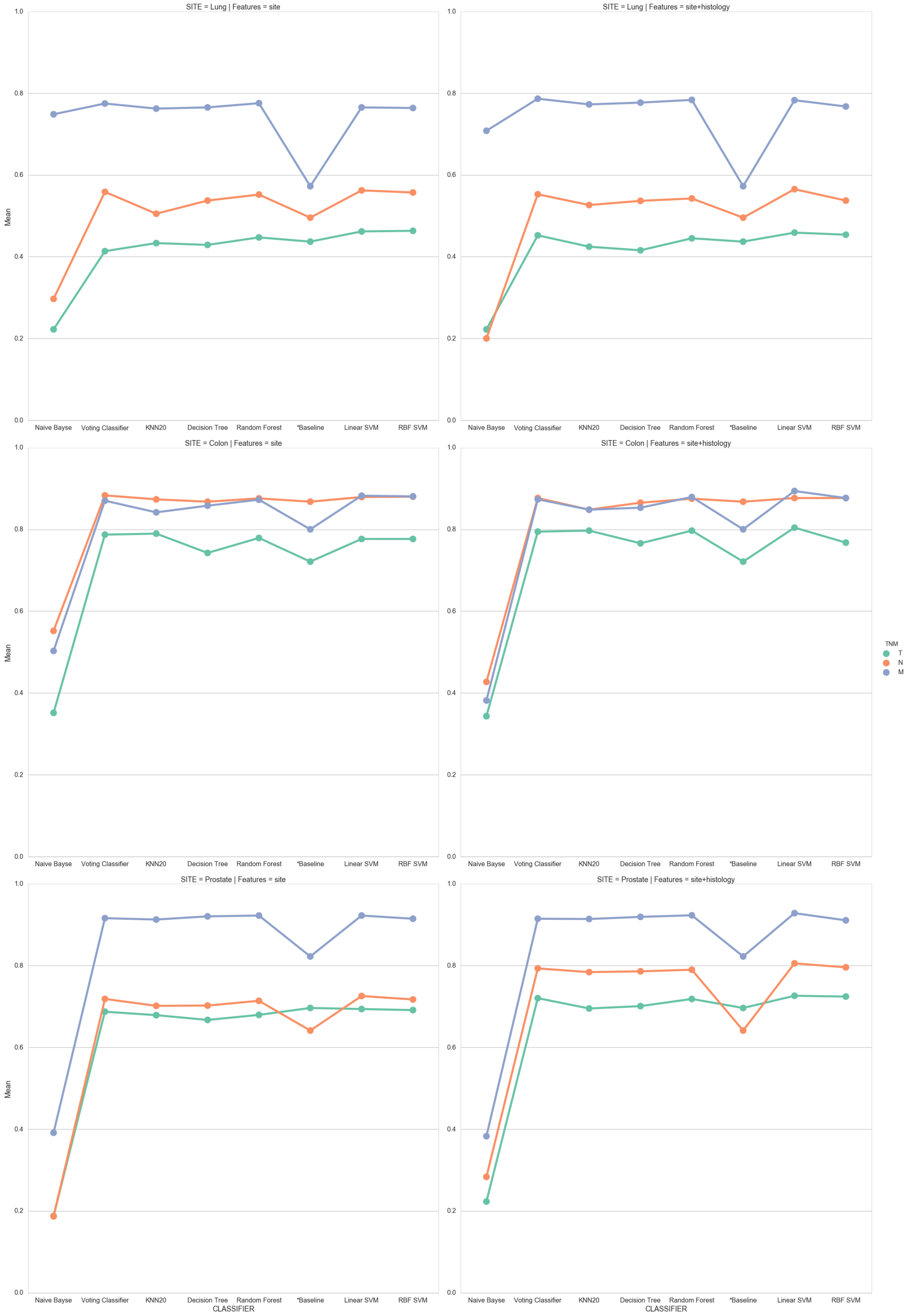


Figure 5.2: The mean accuracy (y-axis) obtained with multiple classification algorithms (x-axis) applied to the training dataset using 3-fold cross validation. From top to bottom, mean accuracy is for all cancer sites, colon, lung and prostate, respectively. Left column is the accuracy obtained when using TNM mentions only while right column is the accuracy obtained when coded histology used as additional feature.

mentions only, and the right column shows accuracy when coded histology is added as a feature. The graph colors distinguish between accuracy scores for T (green), N (orange), and M (blue).

This comparative analysis shows that, in general, accuracy for consolidating T stage is lowest (≈40%–80%) and is best for M stage consolidation (≈80%–90%). This is not surprising given that M is a binary classification problem (M0-M1), and T is a more complex classification problem. The overall best accuracy was obtained with linear SVM and Random Forest classifiers.

Performance with different cancer sites varied for M, with the best performance for prostate (≈90%) and lowest for lung cancer (≈80%), in general, with the exception of performance of T and N which was slightly better for colon (≈80%–90%) than prostate (≈70%-80%) and lung (≈40%–60%).

Adding more features for machine learning improves TNM consolidation accuracy slightly for specific sites. For instance, adding histology feature improves N accuracy for prostate from low 70% to about 80%. Similarly, for colon, adding histology features improves T accuracy slightly from 78% to 80%. This shows that adding more structured data from the unconsolidated NAACCR abstract records could improve accuracy.

5.4.3 Consolidation with Linear Support Vector Machines

The comparative analysis of various machine learning algorithms presented above showed that linear SVM was among the best machine learning algorithms for consolidation of TNM stage information. We chose this algorithm for further analysis, starting with feature selection in which the best combination of structured data items or coded fields (e.g., histology, grade, tumor size) from unconsolidated NAACCR abstracts were selected to maximize classification accuracy. We performed feature selection for colon, lung and prostate cancer sites and each T, N and M stage classification separately. In addition, SVM algorithm[5] has an important hyperparameter known as C value that can be tuned to avoid over-fitting the model to the data and allow for reduced classification error. Using a grid search approach, we found the best C value given the best combination of features found for each site and TNM combinations. Tables 5S.4 and 5S.5 present the mean accuracy and standard deviation obtained when using 3-fold cross validation with the training and development data sets before and after performing feature selections and searching for the best C value. Table 5S.4 presents 3 sets of scores; 1) scores when using extracted TNM mentions only and default value for the C parameter, 2) scores after feature selection (TNM mentions + best features) and 3) scores after performing grid search and using best value for C parameter. Table 5S.5 presents results of the 3-fold cross validation on the 70% training subset and results on 10% development subset obtained with training linear SVM classifier on the 70% training subset.

The results indicate major improvements in accuracy after adding best features for each site especially for the M stage (colon: +9.53%, lung: +16.94%, and prostate: +2.89%). This might be due to the fact that many of the coded fields in NAACCR abstracts capture data associated with metastasis such as nodes involvement (NAACCR data items #820 and 830), metastasis in specific organs (NAACCR data items #2851–2854) and CS site specific factors as part of the collaborative stage system.

Among the three cancer sites, accuracy of T and N stages improved significantly for colon cases when using best features (T: +8.95%, and N: +2.78%). Similarly, there is significant improvement in accuracy for the T and N stages for lung cancer cases (+23.79%, and N: +3.4%).

Performing grid search and using best C value seems to improve performance slightly for colon and lung cancers by an average of 1%. However, there is significant improvement in accuracy for prostate cancer cases by about 6-7% for the T and N stages after using best C value.

5.4.4 Consolidation on the Test Data

The linear SVM classifiers with best features found and C parameter settings were applied to the testing subset after training on the 80% training and development data for each site and TNM combinations. Table 5.3 outlines the results obtained on the held-out testing subset using the Linear SVM for each site and TNM. To perform the weighted kappa, rows that contained TX or TIS for T stage or NX for N stage were removed.

The results show accuracy levels comparable to the results obtained on the cross validation experiments on the training and development subsets. An exception is accuracy for the T stage for both colon and prostate cancer cases which shows a drop of about 5% and a

Table 5.3: Performance obtained on the testing subset with the baseline system and linear SVM trained on the 80% data using best features and C parameters for each site and TNM combinations.

Classifier	Site	TNM	Agreement (%)	Kappa	Weighted
Baseline	(All)	(All)	2358/3567 (66.1%)	0.602	0.984
Baseline	(All)	M	871/1189 (73.3%)	0.229	0.101
Baseline	(All)	N	779/1189 (65.5%)	0.389	0.512
Baseline	(All)	T	708/1189 (59.5%)	0.461	0.616
Baseline	Colon	(All)	810/1062 (76.3%)	0.722	0.988
Baseline	Colon	M	280/354 (79.1%)	0.371	0.154
Baseline	Colon	N	297/354 (83.9%)	0.725	0.843
Baseline	Colon	T	233/354 (65.8%)	0.559	0.740
Baseline	Lung	(All)	593/1182 (50.2%)	0.436	0.951
Baseline	Lung	M	222/394 (56.3%)	0.175	0.132
Baseline	Lung	N	196/394 (49.7%)	0.25	0.329
Baseline	Lung	T	175/394 (44.4%)	0.259	0.359
Baseline	Prostate	(All)	955/1323 (72.2%)	0.655	0.987
Baseline	Prostate	M	369/441 (83.7%)	0.095	0.014
Baseline	Prostate	N	286/441 (64.9%)	0.125	0.400
Baseline	Prostate	T	300/441 (68.0%)	0.519	0.645
Linear SVM	(All)	(All)	2958/3567 (82.9%)	0.803	0.990
Linear SVM	(All)	M	1138/1189 (95.7%)	0.891	0.891
Linear SVM	(All)	N	920/1189 (77.4%)	0.646	0.795
Linear SVM	(All)	T	900/1189 (75.7%)	0.689	0.820
Linear SVM	Colon	(All)	960/1062 (90.4%)	0.888	0.991
Linear SVM	Colon	M	341/354 (96.3%)	0.884	0.884
Linear SVM	Colon	N	323/354 (91.2%)	0.856	0.952
Linear SVM	Colon	T	296/354 (83.6%)	0.785	0.845
Linear SVM	Lung	(All)	888/1182 (75.1%)	0.720	0.980
Linear SVM	Lung	M	370/394 (93.9%)	0.878	0.878
Linear SVM	Lung	N	238/394 (60.4%)	0.435	0.638
Linear SVM	Lung	T	280/394 (71.1%)	0.637	0.808
Linear SVM	Prostate	(All)	1110/1323 (83.9%)	0.803	0.986
Linear SVM	Prostate	M	427/441 (96.8%)	0.812	0.812
Linear SVM	Prostate	N	359/441 (81.4%)	0.609	0.785
Linear SVM	Prostate	T	324/441 (73.5%)	0.618	0.736

drop of about 3% for lung cancer cases. This drop in performance could be partly associated with high standard deviation calculated from the 3-fold cross validation experiments for the T stage which was about 3%.

Figure 5.3 shows the classification matrices for T and N for all sites when compared against the true manually consolidated T and N stages. Most errors seem to center around the diagonal for T stage reflecting the fact that most errors made by the machine were closer

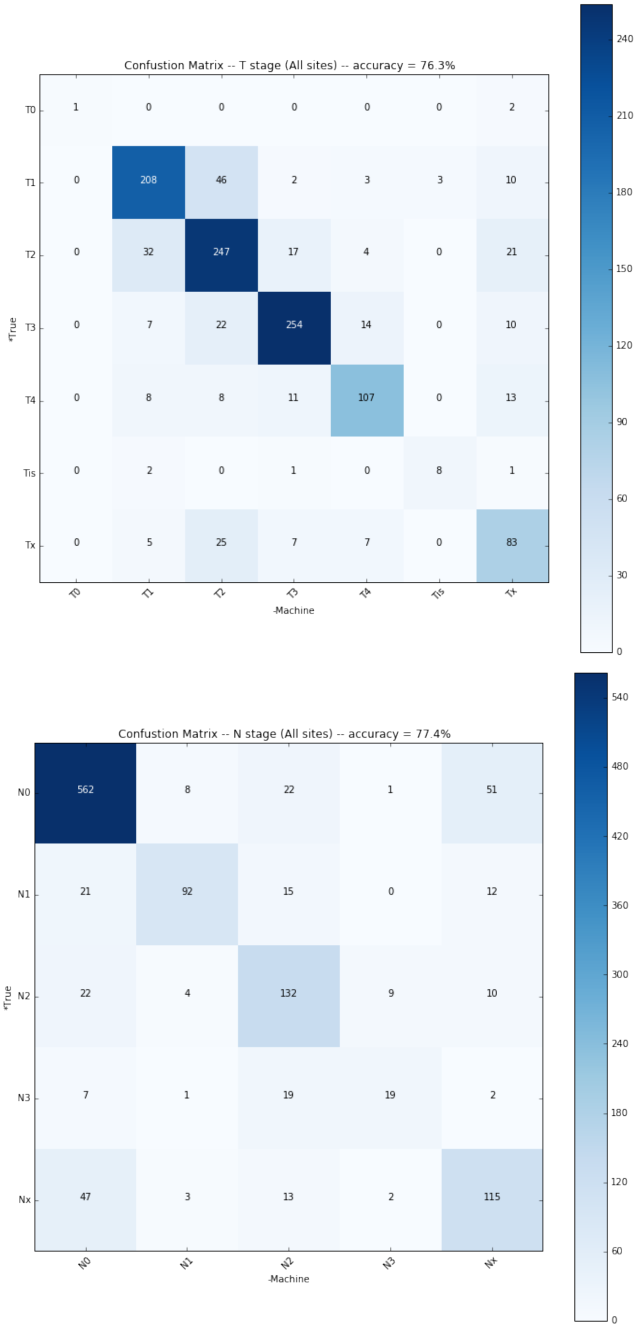


Figure 5.3: Classification errors for T and N stages across all sites on the test set.

to true labels assigned by registrars. Similarly the case for N stage with high proportion of errors (51 and 47) occurring between NX and N0 stages.

5.4.5 Stage Group Derivation

Table 5.4 show the agreement accuracy obtained when deriving the stage group for each cancer case on the testing subset using the TNM predictions from the Baseline and Linear SVM classifiers. Accuracy for colon and lung are above 80% while prostate is lower at 67%.

5.5 Discussion

5.5.1 TNM Consolidation

The cross validation experiments and results on the held-out testing subset using the Linear SVM machine learning algorithm showed that the consolidation of M stage can be reasonably automated with high accuracy (93%-98%). This is a significant improvement on the baseline accuracy for the M stage consolidation (57.29%-82.27%). Across the three sites, consolidation of M stage for lung cancer is lower (93%) than the other two sites considered in this study. This may be due to the prior distribution of the M stage within lung cancer cases where similar proportions of M0 and M1 stages were observed in our dataset, while colon and prostate cases were skewed with far more cases with M0 stages than M1. This is could be one reason the baseline accuracy for the M stage on lung cancer cases is significantly lower than other two sites. Other possible factor that may affect accuracy for the M stage is mentions of MX within the text fields which indicate insufficient information available to stage metastasis. The 7th edition of the AJCC manual eliminated the MX stage and required the use of M0-M1 stage classifications only.

Table 5.4: Performance obtained on the testing subset for derivation of stage group using TNM predictions from baseline and linear SVM classifiers.

TNM Predictions Used	Site	Agreement (%)	Kappa	Weighted
Baseline	(All)	572/1189 (48.1%)	0.353	0.246
Baseline	Colon	231/354 (65.3%)	0.558	0.535
Baseline	Lung	156/394 (39.6%)	0.222	0.113
Baseline	Prostate	185/441 (42.0%)	0.220	0.445
Linear SVM	(All)	942/1189 (79.2%)	0.738	0.875
Linear SVM	Colon	313/354 (88.4%)	0.851	0.928
Linear SVM	Lung	333/394 (84.5%)	0.758	0.883
Linear SVM	Prostate	296/441 (67.1%)	0.584	0.751

Accuracy for consolidation of T and N stages for colon cancer cases was the highest (above 80%) followed by prostate (above 70%) and lowest accuracy for lung cancer cases (above 60%). This is may indicate that the criteria for staging colon cancer cases could be a less complex task than the other two sites. Our study was limited on applying similar techniques regardless of the cancer site to test the level of generalizability of the proposed approach. This approach heavily relies on direct documentation of cancer TNM stage mentions and other coded information in unconsolidated records from hospital registrars. This assumption may not be very realistic and more experiments need to be conducted to potentially improve performance. Most prominent is the use of medical terminologies from the Unified Medical Language System (UMLS) to map medical concepts found in text fields to semantic categories such as anatomical sites, neoplastic processes, and other findings as well as concept codes (CUI). The extracted concepts together with their associated semantic categories and relations between them could then be used in the training process. However, this is may not be a trivial task since there is usually an enormous number of possible mappings and a considerable amount of ambiguity that needs to be resolved. This is especially true with text in the records used for this study which contain extensive use of telegraphic-style sentences, abbreviations, acronyms and all-CAPS text, for instance,

```
LUNG, LUL: 2 CM. MOD TO POORLY DIFF ADENOCA.
```

Knowing that ‘TUMOR’ concept is existing and invades into the ‘ VISCERAL PLEURA’ in the lung could indicate stage T2. Similarly, knowing that ‘8 LYMPH NODES NEGATIVE FOR TUMOR’ could indicate stage N0. Therefore, extraction of concepts and mapping them to UMLS categories and codes may be a useful feature engineering task to improve performance of consolidation in the future. This is may also become necessary when lack of documentation of structured data is an issue for a registry. The rationale behind use of concepts is that patients with similar stages will tend to have similar concepts mentioned in their records and therefore a classifier will learn to associate collections of concepts to a stage. This approach may be promising especially if a registry has large collection of cancer cases records with a sufficient amount of text to allow robust training and classification.

5.5.2 Stage Group Derivation

Deriving a stage group with TNM predictions on the testing subsets showed high accuracy on colon cancer (88.4%) followed by lung cancer (84.5%) while accuracy for prostate cancer was lower (67.1%). Since the AJCC manual specify stage grouping based on TNM subcategories (e.g., T1a-c and N1a-b) and we only predict numerical level TNM and stage groups, we had to compromise when choosing stage grouping in border cases (e.g., stage groups IB–T2aN0M0 and IIA–T1bN0M0 for lung cancer both are derived from T1 numerical stage with different subcategories *a* and *b*). In addition, the criteria for stage grouping for prostate cancer involve the use of PSA and Gleason scores which we do not include in our derivation method and, therefore, the reason for lower accuracy for prostate cases. In future studies, we propose to proceed in two phases to handle the extra layer of complexity involving use of alphabetical subcategories within the labels. In the first phase, group TNM labels into a coarse-grained numerical level and then build classifiers to predict numerical stage only. Second, use results of the previous step as input to new classifiers with original information that can predict the more refined alphabetical stages in a cascading fashion. This process will allow measuring performance of classification at the two levels separately and could potentially reduce label complexity by dealing with a smaller number of categories in each classification task.

5.6 Conclusion

Automatic consolidation of cancer stages by machine learning for the cancer registry could achieve high accuracy for some cancer sites and may be practical and useful in the context of manual human review assistance. Future studies may focus on additional feature engineering and extraction of medical concepts to improve consolidation accuracy.

5.7 Contributors

AKA is the main developer of the application and prepared initial version of the manuscript. SM provided valuable contributions to the study design and made corrections/additions to the manuscript. All other authors have participated in the study and read/edited the manuscript.

5.8 Funding

This work has been supported by research contract number HHSN261201300017I from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) program.

5.9 Competing Interests

None.

5.10 Supplementary Materials

This section contains additional tables that can be included as supplementary materials for Chapter 5.

Table 5S.1: Distribution of cases by set assignment and cancer site

	Lung	Colon	Prostate	Total
Training	1365	1228	1540	4133
Validation	194	178	221	593
Testing	394	354	441	1189
Total	1953	1760	2202	5915

Table 5S.2: Example stage group classification based on TNM stages for the primary cancer sites considered in the study according to the AJCC manual, 7th edition.

Cancer Site	Stage Group	T	N	M	PSA	Gleason
Prostate	I	T1a-c	N0	M0	PSA $<$ 10	Gleason $\leq$ 6
	I	T2a	N0	M0	PSA $<$ 10	Gleason $\leq$ 6
	I	T1-2a	N0	M0	PSA X	Gleason X
Colon	IIA	T3	N0	M0		
	IIB	T4a	N0	M0		
	IIC	T4b	N0	M0		
Lung	IIA	T2b	N0	M0		
	IIA	T1a	N1	M0		
	IIA	T1b	N1	M0		
	IIA	T2a	N1	M0		
	IIB	T2b	N1	M0		
	IIB	T3	N0	M0		

Table 5S.3: Accuracy scores (mean and standard deviations from 3-fold) obtained with the baseline approach on the 70% training data.

TNM	Cancer Site	Mean Score	Stdev
T	All sites	61.84%	1.10%
N	All sites	66.07%	1.11%
M	All sites	73.36%	2.51%
T	Lung	43.73%	2.76%
N	Lung	49.59%	3.68%
M	Lung	57.28%	2.98%
T	Colon	72.15%	1.88%
N	Colon	86.80%	0.58%
M	Colon	80.05%	2.28%
T	Prostate	69.67%	4.23%
N	Prostate	64.15%	3.69%
M	Prostate	82.27%	1.99%

Table 5S.4: Accuracy scores (mean and standard deviations from 3-fold cross validation) obtained when performing feature selection and C parameter search for linear SVM algorithm on the 80% combined training and development subsets.

Features	TNM	Cancer Site	Mean Score	Stdev	C value
TNM mentions	T	Colon	78.03%	2.38%	1
TNM mentions	N	Colon	87.97%	0.53%	1
TNM mentions	M	Colon	88.26%	1.37%	1
TNM mentions	T	Lung	47.60%	3.49%	1
TNM mentions	N	Lung	55.48%	2.73%	1
TNM mentions	M	Lung	76.90%	0.81%	1
TNM mentions	T	Prostate	69.61%	0.53%	1
TNM mentions	N	Prostate	72.23%	1.68%	1
TNM mentions	M	Prostate	92.22%	1.10%	1
TNM mentions + best features	T	Colon	86.98%	0.53%	1
TNM mentions + best features	N	Colon	90.75%	1.34%	1
TNM mentions + best features	M	Colon	97.79%	0.36%	1
TNM mentions + best features	T	Lung	71.39%	1.99%	1
TNM mentions + best features	N	Lung	58.88%	1.07%	1
TNM mentions + best features	M	Lung	93.84%	1.96%	1
TNM mentions + best features	T	Prostate	72.84%	3.10%	1
TNM mentions + best features	N	Prostate	74.95%	1.25%	1
TNM mentions + best features	M	Prostate	95.11%	1.61%	1
TNM mentions + best features	T	Colon	88.40%	0.74%	0.12
TNM mentions + best features	N	Colon	91.82%	0.28%	0.075
TNM mentions + best features	M	Colon	98.15%	0.09%	0.045
TNM mentions + best features	T	Lung	74.53%	3.18%	0.115
TNM mentions + best features	N	Lung	59.39%	0.69%	0.135
TNM mentions + best features	M	Lung	94.41%	1.96%	0.06
TNM mentions + best features	T	Prostate	78.64%	3.09%	0.02
TNM mentions + best features	N	Prostate	81.37%	1.48%	0.025
TNM mentions + best features	M	Prostate	97.61%	0.14%	0.01

Table 5S.5: Accuracy scores on the 70% training subset and 10% development using best features and C value found for linear SVM. Training score: mean accuracy from 3-fold cross validation on 70% training subset. Development score: accuracy obtained on the 10% development subset after applying classifier trained on the 70% training subset.

TNM	Cancer Site	Training Score	Stdev	Development Score
T	Colon	87.69%	0.80%	88.20%
N	Colon	91.12%	0.59%	89.32%
M	Colon	97.96%	0.30%	98.31%
T	Lung	71.86%	1.23%	78.86%
N	Lung	58.60%	1.31%	60.30%
M	Lung	93.70%	1.28%	93.81%
T	Prostate	78.23%	3.22%	77.82%
N	Prostate	81.62%	1.15%	80.99%
M	Prostate	97.66%	0.82%	98.19%

5.11 References

[1] A. K. AAlAbdulsalam, J. H. Garvin, A. Redd, M. E. Carter, C. Sweeny, and S. M. Meystre, *Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry*, in AMIA 2018 Informatics Summit, March, 2018 (Submitted).

[2] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen, *Semeval-2016 task 12: Clinical tempeval*, Proceedings of SemEval, (2016), pp. 1052–1062.

[3] K. D. Faber, V. K. Cortessis, and S. Daneshmand, *Validation of Surveillance, Epidemiology, and End Results TNM staging for testicular germ cell tumor*, in Urologic Oncology: Seminars and Original Investigations, vol. 32, Elsevier, 2014, pp. 1341–1346.

[4] K. Herget, A. Stroup, K. Smith, M. Wen, and C. Sweeney, *Unstaged cancer: Long-term decline in incidence by site and by demographic and socioeconomic characteristics*, Cancer Causes & Control, 28 (2017), pp. 341–349.

[5] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al., *A practical guide to support vector classification*. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, 2003.

[6] A. C. Klassen, F. Curriero, M. Kulldorff, A. J. Alberg, E. A. Platz, and S. T. Neloms, *Missing stage and grade in Maryland prostate cancer surveillance data, 1992–1997*, American Journal of Preventive Medicine, 30 (2006), pp. S77–S87.

[7] W.-L. Liu, S. Kasl, J. T. Flannery, A. Lindo, and R. Dubrow, *The accuracy of prostate cancer staging in a population-based tumor registry and its impact on the black-white stage difference*, Cancer Causes and Control, 6 (1995), pp. 425–430.

[8] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, et al., *Extracting information from textual documents in the electronic health record: A review of recent research*, Yearb Med Inform, 35 (2008), p. 44.

[9] P. Minicozzi, K. Innos, M.-J. Sánchez, A. Trama, P. M. Walsh, R. Marcos-Gragera, N. Dimitrova, L. Botta, O. Visser, S. Rossi, et al., *Quality analysis of population-based information on cancer stage at diagnosis across europe, with presentation of stage-specific cancer survival estimates: A eurocare-5 study*, European Journal of Cancer, 84 (2017), pp. 335–353.

[10] D. M. Parkin, *The evolution of the population-based cancer registry*, Nature Reviews Cancer, 6 (2006), p. 603.

[11] L. A. Ries, D. Harkins, M. Krapcho, A. Mariotto, B. A. Miller, E. J. Feuer, L. X. Clegg, M. Eisner, M.-J. Horner, N. Howlader, et al., *SEER Cancer Statistics Review, 1975-2003*, (2006).

[12] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, *Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications*, Journal of the American Medical Informatics Association, 17 (2010), pp. 507–513.

[13] S. Seneviratne, I. Campbell, N. Scott, R. Shirley, T. Peni, and R. Lawrenson, *Accuracy and completeness of the New Zealand cancer registry for staging of invasive breast cancer*, Cancer Epidemiology, 38 (2014), pp. 638–644.

[14] I. Spasić, J. Livsey, J. A. Keane, and G. Nenadić, *Text mining of cancer-related information: Review of current status and future directions*, International Journal of Medical Informatics, 83 (2014), pp. 605–623.

[15] A. Stewart, A. Hurlbut, L. Havener, F. Michaud, S. Capron, L. Ries, et al., *North American Association of Central Cancer Registries NAACCR 2006 implementation guidelines and recommendations*. `http://www.naaccr.org/`, 2012.

[16] W. Sun, A. Rumshisky, and O. Uzuner, *Evaluating temporal relations in clinical text: 2012 i2b2 challenge*, Journal of the American Medical Informatics Association, 20 (2013), pp. 806–813.

[17] M. Thornton and L. OConnor, *Standards for cancer registries volume ii: Data standards and data dictionary, record layout version 12.2*, North American Association of Central Cancer Registries, (2012).

CHAPTER 6

DISCUSSION

6.1 Summary

The ultimate goal of this study is to demonstrate the utility of NLP and ML to automate cancer stage consolidation in a central cancer registry. I have presented a preliminary study that involved reusing and adapting exiting NLP resources and medical terminology to solve new clinical information extraction task. Reusing existing NLP resources provided very good baseline and reduced development efforts. I then compared performance of two well known statistical machine learning algorithms (CRF and SVM) in extracting cancer-related concepts, time expressions and relations from clinical text of colon cancer patients. The evaluation showed that while both algorithms performed comparably in general, CRF performed slightly better for sequential extraction; in addition, combining both CRF and SVM in ensemble-based approach may improve performance for time expressions extraction.

Automatic extraction of TNM stage mentions achieved very high performance suggesting usefulness of the proposed NLP system based on simple regular expressions in future tasks. There were aspects about TNM mentions that were not clear at the start of the project, in particular, the negation, historical and temporal contexts for these kinds of mentions. These contexts became clear after the completion of the reference standard development which was necessary to evaluate the NLP system and better understand the information context surrounding these TNM mentions. The TNM mentions were mostly affirmed, recent, and related to the patient in question, and did not occur in very sophisticated contexts. This is to large extent indicated that the proposed NLP system likely benefited from having these TNM mentions occur in direct contexts and made development easier since the amount of context that needs to be extracted and analyzed is minimal.

TNM stages could be either clinical or pathological based on general criteria involving the time frame of staging and the use of clinical exams, imaging tests and biopsies or surgical resection of tumor and pathological examination to determine stage. Knowledge of clinical and pathological stages is imperative for effective treatment decisions and future analysis of registry data. The proposed CRF approach to classify TNM stages to clinical or pathological achieved high sensitivity (83.5%–87%) significantly better than a baseline rule-based system (38.4%–53.5%).

The extracted TNM stage mentions and use of other structured data fields from patient records to automatically consolidate TNM stages at the patient level using machine learning proved to be feasible based on results obtained from the cross validation experiments and on the testing set. Automatic consolidation could achieve very high accuracy for the M stage for all three cancer sites considered in the study (above 90%), very good accuracy for the T and N stages for colon cancer (above 80%), moderate accuracy for prostate cancer (70%–80%) and lower accuracy for lung cancer (60%–70%). The proposed machine learning algorithms do not make assumptions about underlying criteria to stage cancer cases or do not attempt to capture information criteria that may allow direct staging such as tumor size or number of lymph nodes involvement. The criteria for staging cancers are complex and differ for each site and algorithms that attempt to extract information which can be used to infer stages will be difficult to generalize without retraining them with new data for each different site. This will mean the requirement to develop a new annotation schema and reference standard for each different site detailing the data items required for staging, a very expensive and demanding task. The proposed algorithms is more pragmatic in the sense that we try to depend on existing TNM stage documentation and other coded data from reporting facilities to consolidate stages for central cancer registry.

Deriving a stage group based on the automatically consolidated TNM stages showed promising results for Colon and Lung cancers while deriving a stage group for Prostate cancer attained lower accuracy because both gleason and PSA score values which are required for deriving a stage group were not extracted by the proposed NLP system.

6.2 Future Directions

The nature of clinical text collected at the Utah central cancer registry varies mainly because these text fields are collected from various reporting facilities ranging from large regional hospitals to small community health centers. In addition, the text is a verbatim copy of notes from relevant section of the patients EHR or manual dictations by local cancer registrars. The most prominent feature of this text is the extensive use of uppercase letters which makes it difficult to easily discern abbreviations, and distinguish proper nouns. In addition, the use of short telegraphic style sentences, abbreviations and acronyms is very common such as “COULD NOT FIND EVIDENCE OF ANY FURTHER TX” or “CC: AB PAIN”. This could be challenging for NLP systems and novel approaches may be needed to handle the extensive use of abbreviations and all-caps text.

Each patient case in the registry has multiple records with text fields that were used to extract TNM mentions. The TNM mentions together with other coded data were consolidated using the proposed SVM algorithm for each cancer site at the patient level. Although this approach yielded promising results, it is likely that more feature engineering or structuring of the prediction task will be required to enhance performance. NLP tasks usually assume a document structure with a class label assigned at either document-level (for document classification) or word-level (for information extraction). This project involved eventually assigning a label at the patient-level (i.e., a label for multiple documents). Since each patient may have multiple records with text fields (and TNM mentions), the task can be structured such that labels are assigned at the document-level for each separate record and then using some heuristics to assign final label at the patient-level based on previously predicted document-level labels. Alternatively, the multiple records per patient may be combined using temporal labels and then using some heuristics to discard duplicated fields. This latter approach will likely require sophisticated feature engineering to find most salient features for final stage consolidation at the patient-level. This is also a venue for experimenting with more novel feature representations for this task.

Another challenge for the proposed consolidation algorithms is the lack of adequate documentation of TNM stage mentions required to consolidate a final stage. This may be mitigated by the use of bag-of-concepts or more refined concept-level representations. The rationale behind this is that cancer cases with similar TNM stages might tend to have similar mentions of concepts. For instance, colon cancer cases that have mentions of affirmed 'TUMOR' that penetrates the 'VISCERAL PERITONEUM' concepts will more likely be staged T4. Similarly, these concepts may be mentions of tumor invasion within specific anatomical sites such as penetration of 'TUMOR' cells into the 'SUBMUCOSA' for colon cancer (mostly indicating T1 stage), or the mention of 'DISTANT METS' to mostly indicate M1 stage for most cancers. However, as mentioned previously the extraction and mapping of concept terms and relations between them may not be a trivial task given the nature of the text. It is likely that a considerable amount of ambiguity needs to be resolved before a useful representation maybe used to train a module. The ambiguity could be managed by the use of more recent word embedding and deep learning to build word vectors that capture semantics at a deeper level. This could be a viable direction to pursue especially if a large amount of textual data becomes accessible in the future (i.e., millions of words).

Consolidation of T, N and M stages was performed separately by developing SVM classifiers for each site and TNM combinations. It is likely that future implementation may benefit from experimenting with, for instance, feeding T and N classifiers the predictions of M stages or vice versa. The rationale behind this is the strong likelihood of advanced T stage (T2–T4) given previously predicted metastasis stage (M1) for a cancer case. Therefore, each classifier will inform the other classifiers in a feedback loop that could potentially improve performance.