2013 IEEE 13th International Conference on Data Mining Workshops

MedCat: A Framework for High Level Conceptualization of Medical Notes

Samah Jamal Fodeh MS, PhD
Yale University School of Medicine
Connicticut, USA
Email: samah.fodeh@yale.edu

Maryan Zirkle MD, MS, MA
Portland Veterans Affairs Medical Center
Oregon, USA
Email: maryan.zirkle@va.gov

Dezon Finch PhD
James A.Haley Veterans Affairs Medical Center
Florida, USA
Email: dezon.finch@va.gov

Cynthia Brandt MD, MPH
Joseph Erdos MD, PhD
Veterans Affairs Connicticut Helathcare System
Connicticut, USA
Email:cynthia.brandt@va.gov,
joseph.erdosMD@va.gov

Ruth Reeves PhD
Tennessee Valley Health System - Department of Veterans Affairs, Tennessee, USA
Email: ruth.reeves2@va.gov

Abstract— **In this paper we introduce a new framework called MedCat to delineate and demonstrate an approach for projecting representations of concept-derived content in clinical notes into a new categorization space to reduce dimensionality and noise in the data. Constructing MedCat framework required several steps including manual annotation, knowledge base expansion using MetaMap, concept category construction, automated annotation using NLP to generate a bag of concepts, and finally concept conversion to higher level abstracted categories. The framework was applied to Post Traumatic Stress Disorder (PTSD) clinical notes for evaluation. A random sample of PTSD clinical note content was automatically recategorized into six PTSD treatment categories using MedCat. Using existing annotations from PTSD notes that were categorized by content experts into treatment categories as the reference standard, the sensitivity of the framework in detecting the treatment categories was greater than 90%. The results suggest that representations of concept-derived content when categorized by relevance features can be used to reliably understand and summarize clinical notes.**

Keywords-component; MedCat, PTSD, Categorization, Conceptualization, Natural Language Processing

I. INTRODUCTION

Useful clinical information is included in narrative form in clinical documents (i.e. discharge summaries and progress notes) in electronic health records (EHRs). While this textual data might allow for rich description of important clinical information, a key challenge is that relatively few guidelines are given to providers on how and what to record and this, along with different documentation styles, results in inconsistency and variation in documentation of important clinical information. This variation introduces noise and many other challenges to methods proposed to automatically process narrative documentation for extracting clinically useful information. There are, for example, methods involving Natural Language Processing (NLP) techniques [18],[5],[8],[15],[17] that parse complex narrative text and describe the text using different components such as token words, sentences or clinical concepts from the Unified Medical Language System UMLS [23]. In this study, we propose a new framework that creates a representation of the concepts within the clinical notes in a new space with higher levels of abstraction. This abstracted conceptual layer can facilitate readability and interpretation of notes as it can reduce dimensionality and noise in the data. In this framework the output of mapping phrases within clinical notes to UMLS concepts collapses mapped concepts into more general levels or categories. The gain from the new data representation is the reduction of the dimensionality i.e. the size of the feature set and elimination of the effect of irrelevant concepts (noise). We demonstrate the utility of MedCat on clinical notes to extract different types of psychotherapeutic treatment modalities for patients with PTSD. The major contribution of this paper is the description of a method to reduce noise by building a new representation of data based on categorizing the relevant concepts at a higher level.

II. RELATED WORK

Many NLP and machine learning-based techniques have been proposed to parse through large volumes of unstructured clinical narratives to extract clinically useful information. Thus we focus on NLP systems that map clinical narrative text to ontologies. Generalized language open-source frameworks such as the Generalized

978-0-7695-5109-8 2013

DOI 10.1109/ICDMW.2013.89

Architecture for Text Engineering (GATE) [4] and the Unstructured Information Management Architecture (UIMA) [5] enabled medical informatics developers to build task-specific modules or create general clinical NLP pipelines such as the Clinical Text Analysis and Knowledge Extraction System (cTAKES) [15] and the Yale CTAKES extension for document classification YTEX [10]. Other examples of NLP systems for mapping clinical free-text to concepts include MedLee [10], KnowledgeMap [6], ONYX [2], MedEX [24], and MCVS [7]. Different feature extraction processes such as word tokenization, UMLS concept identification and sentence splitting constitute the production stream of an NLP pipeline. These features are subsequently combined with supervised machine learning techniques to carry out specific inferencing tasks. The strength of these techniques depends on the features that they use. Existing methods often use UMLS concepts as features and combine them with syntactic features or grammatical features to improve performance [20]-[22],[25]. In our previous work [17] and in the work of Zeng et al [14] a concept's feature set was augmented by categorical information of the concepts represented by their semantic types in UMLS. We noticed that this categorical information noticeably improved the results. However, the semantic types in UMLS are very general; for example all treatment concepts belong to a single treatment semantic type. This lack of granularity causes loss of information and does not add much value to the feature set. This is particularly problematic when the inferencing task requires differing categories to be recovered from the data, such as pharmacology and psychotherapeutic treatments. In this article we investigate this matter in more depth and show that by customizing the features and replacing them with their subsuming concepts or categories within a predefined concept-category hierarchy, we reduce the size of the feature set, capture the high level context and increase the readability of the clinical notes.

As reported by Stanfill et al [20] after reviewing more than 100 articles of automated coding and classification systems, they concluded that most of the existing systems are not generalizable and only serve a specific purpose. Unlike these methods and because our framework summarizes the clinical text to facilitate readability using a learned hierarchy of concepts, it can be generalized to solve different problems once the concept-category hierarchy is defined.

III. METHODS

This section describes how the dataset was collected and outlines the development steps of our framework for generating a new categorized description of clinical notes to identify the different psychotherapeutic PTSD treatment types documented. The performance of our framework is evaluated by the comparison of the types of treatments that were automatically extracted in each note using NLP against the treatment types assigned by clinically trained annotators.

A. Dataset

The Electronic Medical Record (EMR) documents comprised of clinical notes for this study were drawn from a cohort of VA patients with at least two outpatient visits between 10/1/2010 and 9/30/2011 with a primary coded diagnosis (ICD-9) for PTSD. A subset of these text notes were then selected based on the Department of Veterans Health Affairs (VHA) Enterprise Standard Title with the phrase "Mental Health" in the title. The selected notes represented patients from 125 VA facilities, representing more than 140 hospitals. A random sample from these notes was drawn and annotation sets were made in groups of approximately 100 notes, for a total of 585 clinical notes in the study corpus. The truth labels or reference standard is comprised of manual annotations done at the Portland VA Medical Center by 2 annotators with disagreements resolved by a third content expert adjudicator. The adjudicated results formed the reference standard used for evaluating the performance of our framework. A detailed PTSD guideline and schema was developed to cover specific details related to treatments of PTSD.

B. PTSD Knowledge Base (PTSD-KB)

For this project, our purpose was to augment a knowledge base and use it in the application of NLP tools to further the discovery process and make inferences about a specific medical domain - PTSD. To do this, first we needed to define the vocabulary used by mental healthcare providers to describe the clinical course of our intended target domain: PTSD. In order to provide our NLP system with the foundation upon which to build a concept extraction system, optimized for conceptual relevancy, a highly focused vocabulary was required. The concepts and relationships were derived from other terminology (specifically, SNOMED-CT), clinical guidelines, focus groups, cognitive interviews with a variety of mental health providers and annotation of clinical documents. A total of 370 treatment terms were added to the PTSD-KB from the manual annotation process. The final PTSD-KB contained several treatment terms including psychotherapeutic techniques and modalities not documented in any other comprehensive terminology.

C. Framework for High Level Clinical Data Conceptualization

The components of MedCat are shown in Figure 1 and consist of several steps. It starts with manual annotation and knowledgebase expansion (*PTSD-KB)* using MetaMap, followed by concept-category hierarchy construction, then automated annotation to generate bag of concepts and finally transformation of concepts into bag of categories that represents the higher level abstraction of the concepts identified in clinical notes. Figure 1 displays the details of MedCat framework.

- Manual Annotation: The manual annotation was done using Knowtator [13]. Knowtator is a general-purpose text annotation tool that is integrated with the Protégé knowledge representation system. Knowtator facilitates the manual creation of training datasets for biomedical language processing purposes. Because the test bed

application is PTSD, basic annotation guidelines related to PTSD treatment were provided and the creation of more defined and directed guidelines was an iterative process during the ongoing annotation efforts. The annotation task was concerned with terminology used to express different types of treatment documented for patients with PTSD. This effort also captures any documentation of non-compliance from a patient for a prescribed treatment and subsequent negation terms. A total of 370 treatment terms were added to the PTSD-KB after the manual annotation process to be used in automated annotation.

- Knowledge Base Expansion Using MetaMap: To avoid over-fitting the system to the data, expanding the PTSD-KB to a wider coverage was critical. To this end, we exploited the functionalities of MetaMap. MetaMap is a tool developed at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text.

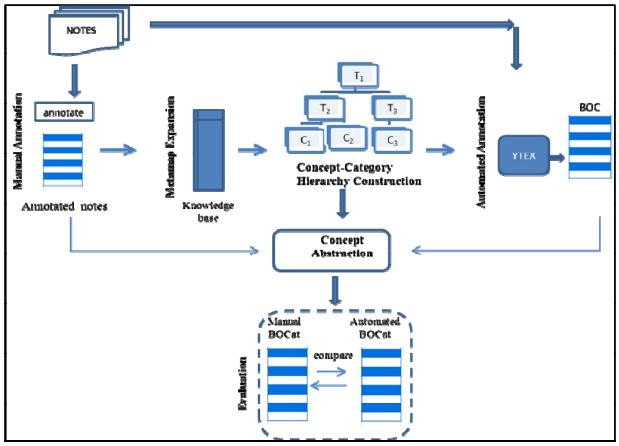

Figure 1: Customized data representation Framework MedCat

MetaMap uses a knowledge-intensive approach based on symbolic, natural language processing (NLP) and computational-linguistic techniques [1]. In order to expand the *PTSD-KB* with synonyms, acronym definitions and other related concepts from UMLS, we loaded the annotations (text spans of PTSD notes and their classifications) that were tagged by the human annotation team into MetaMap. We parsed the XML output of MetaMap and extracted 426 concepts and added these to the *PTSD-KB,* which originally contained only 370 terms. After resolving overlaps between the original PTSD-KB concepts and the MetaMap concept output, the size of the *PTSD-KB* was equivalent to 645 concepts.

- Concept-Category Hierarchy Construction: Drawing a more summarized version of the data is important to comprehend and analyze different clinical problems in clinical notes. For PTSD in particular, the categorization of treatment information in clinical notes is necessary to assess the effect of different types of treatments compared to the clinical status of the patients. In MedCat, we categorize the information in notes by mapping the extracted units of analysis (concepts) from YTEX to more general categories. In order to achieve this goal, we abstracted the vocabulary in the knowledge base into a concept-category hierarchy of conceptual granularity such that the most specific concepts are placed at the bottom of the hierarchy and the more general concepts are placed further up. The complexity of the structure and the semantic network of the hierarchy will vary relative to the application domain and research question. A hierarchy may include many levels. In our application, we constructed a simple 2-level hierarchy with one type of hierarchical relationship called "part_of". While the bottom level of the concept-category hierarchy is comprised of the detailed treatment concepts mentioned in the text note, the upper level encodes the treatment categories of these concepts. An even higher level could have used the UMLS semantic types such as “Treatments, SemanticType “STY | T121 | Pharmacologic Substance semantic type” but we found these to be too general for our purpose. We identified five psychotherapeutic categories into which the concepts in the *PTSD-KB* were classified: Pharmacotherapy, Psychological, Psycho-Social, Psycho-Education, and Case Management. A category “Other” is also created to capture any terms that were extracted, but did not fit into any of the other five categories. TABLE I shows a sample of treatment terms extracted from the text along with its mapped category. Some of the concepts produced by MetaMap did not have an obvious treatment category, such as “control function, comprehension and component”. Those concepts were given the type 'non relevant'.

TABLE I: PTSD TREATMENT CATEGORIES AND RELATED EXAMPLE TERMS

Pharma-cology	**Psycho-logical**	**Psycho-Social**	**Psycho-Education**	**Case Manage-ment**	**Other**
Wellbutrin SR	Group Therapy	Recreational activity	Pamphlet	Connection Claim	Admiss-ion
Propranolol	Empathy	Communication	Crisis manage-ment	Encouraged to Speak to Psychiatrist	Admit to
Zolpidem	Rapport	Social contact	Stress manage-ment	Case manage-ment (procedure)	Chapel services

Because MetaMap breaks a phrase (phrases of manual annotations in our case) into individual words then maps each word to UMLS to check for existence or synonyms, many of the output concepts of MetaMap are not necessarily related to treatments but were included in the output because their first word matched the mapped word.

- Automated Annotation: We used YTEX [10] to automatically annotate the clinical notes. YTEX is an extension to the clinical Text Analysis and Knowledge Extraction System (cTAKES) to derive robust feature sets from NLP pipelines [1]. Different types of features are generated by YTEX such as words, concepts, phrases or sentences. Our interest was to extract concepts only and use them as features. Concepts from UMLS are stored in a YTEX dictionary and are used by the named entity recognition module to annotate concepts. Since our focus in this application is to identify PTSD treatment related concepts only and because UMLS includes different types of concepts from a variety of sources, we replaced the built-in dictionary of YTEX with the *PTSD-KB* trained by the combination of manual annotations of PTSD notes and metaMap mappings. This dictionary customization resulted in a reduced set of features (concepts) that we subsequently extracted from YTEX output to represent the notes. We call this new representation of the data Bag Of Concepts (BOC). More specifically, the BOC is a matrix where the rows are the PTSD treatment notes and the columns are the treatment concepts.
- Concept Abstraction - BOC to BOCat Transformation: Because the BOC extracted from YTEX contains a variety of features, some of which are relevant to our interests and others not, those features must be processed somehow to better serve the desired purpose. We present such a process to transform the detailed features (concepts) extracted from the notes into more general i.e. less granular set of features by integrating background knowledge from a specialized concept-category hierarchy. That is transforming the representation of the text notes from concepts to a more abstracted feature space of categories. The benefit of this transformation is three-fold: first it reduces the complexity of further analysis by decreasing the dimensionality of the space from hundreds of concepts to the size of the new abstract space. Second, it enhances the readability of the notes by removing the redundant features that are not relevant and only obscure the analysis. Third, it may signify different categories in the new space to capture and conceptualize the data for better understanding. We transformed the BOC representation of the PTSD clinical notes to the Bag Of Categories (BOCat) representation, where the categories are the types of PTSD treatments. A huge reduction of dimensionality is achieved using BOCat. In the BOC representation, we had 645 concepts to represent the notes whereas in the BOCat representation, the feature space is compressed and the notes are described using only 6 dimensions representing the categories of PTSD treatments. In addition, mapping the concepts to their corresponding categories using the concept-category hierarchy shifts focus to only those concepts. It is important to mention that *PTSD-KB* is comprised of a variety of concepts including those that are not related to treatments and resulted from the automated mapping of annotations in MetaMap as well as the manual annotations which usually encompasses human error. YTEX is anticipated to propagate this error or noise (because such concepts are included in its dictionary) thereby generate concepts unrelated to PTSD treatments. To generate the BOCat without the noise of irrelevant concepts, we map each concept to its treatment type using the learned concept-category hierarchy. We thus fashioned a filter, wherein concepts that were mapped to the type 'non relevant' are dropped from the analysis. We assigned this type to concepts that were not related to actual treatments. In addition to noise and dimensionality reduction, the BOCat representation gives weights to different types of treatments in the notes. The weight of a treatment type is calculated in a particular clinical note by adding up the frequencies of all concepts belonging to that type of treatment. This information, typically documented exclusively in the narrative text, indicates how often a treatment type is documented in a clinical note and allows for the information to be compared to a patient's existing PTSD symptoms.

IV. EVALUATION

To evaluate our framework in generating a more concise description of PTSD notes using treatment types, we used the treatment type annotations that we entered manually using Knowtator as reference standard. To facilitate the evaluation, we built a BOCat representation (manual_ BOCat) based on manual annotation similar to what we generated from the automated annotations (automated_ BOCat). To compare both manual_BOCat and automated_ BOCat, we define two variables: ***Undetected*** and ***Detected***. The variable ***Undetected*** measures the number of notes where MedCat failed to identify at least one treatment. The variable ***Detected*** measures the number of notes in which treatment types were extracted and correctly mapped to their categories. To measure the overall performance of MedCat, we aggregated the results in a contingency table as shown in TABLE II. The statistics in the table represent the following information:

a- number of notes in which MedCat has detected and failed to detect treatment types in these notes.

b- number of notes in which MedCat completely identified treatment types.

c- number of notes in which MedCat failed to detect any treatment type.

d- number of notes in which MedCat did not detect or miss any treatment type - this cell in the table is always equal to zero since it does not apply to any meaningful event as the framework should either detect or fail to detect treatment types.

The best performance of MedCat occurs when the value of (b) is equal to the total number of treatment notes. To measure the ability of the framework to detect each treatment type individually, we computed the sensitivity of each treatment type. For each treatment type *t* we divided the total number of notes that contain *t* according to the automated_BOCat data representation by the total number of notes that have *t* according to manual_BOCat.

TABLE II: CONTINGENCY TABLE FOR EVALUTION

		Undetected	
		Yes	*No*
Detected	*Yes*	a	b
	No	c	d

V. RESULTS

Once the manual annotation was completed, we processed the output from Knowtator coded in XML files to extract treatment annotations. Out of the 585 clinical notes annotated with specific details related to treatments of PTSD, 162 clinical notes were actual treatment notes because they included PTSD treatment terms. The remaining notes had other annotations some of which related to symptoms and other mental health or administrative terms. Because our goal is to identify treatment types in clinical notes, we focused only on the 162 treatment notes and had our framework automatically identify their contained treatment concepts and the associated treatment types or categories. We first generated the manual_BOCat matrix from the manual annotations. The matrix had 162 rows corresponding to the treatment notes and six columns of treatment types as follows: 1 Pharmacotherapy, 2 Psychological, 3 Psycho-Social, 4 Psycho-Education, 5 Case Management and 6 others. We generated the automated_BOCat from the automated annotations produced by YTEX. The results (TABLE III) suggest that MedCat completely detected treatments in 136 out of 162 treatment notes, at a 91% rate. The false negatives (FN) rate was 9% where MedCat missed treatment categories in 15 notes. We found that in 11 notes (6%) some treatment types were identified and some were missed.

TABLE III: DETECTED VERSUS UNDETECTED TREATMENTS IN PTSD NOTES

		Undetected	
		Yes	*No*
Detected	*Yes*	11	136
	No	15	0

We have also investigated the sensitivity of finding individual treatment types. The results are displayed in TABLE IV and Figure 2. The first row labeled “Manual_BOCat” displays all treatment types based on the manual annotations. The sensitivity of each treatment type (shown in the 6 columns) is computed by dividing its value by the corresponding value in the Manual_BOCat row. Pharmacology, Psychological, and Psycho- Education were detected at high sensitivities respectively: 90.7%, 89% and 81%. We expect Pharmacology type of treatment to have the highest sensitivity since the pharmacology terminology is well defined in *PTSD-KB* in particular and in medicine in general.

We examined the notes corresponding to false negatives to understand why the framework could not detect treatment mentions in these documents. The detection failure could be attributed to three reasons: first, treatments in some of these notes were indeed annotated by YTEX but were negated in the context thereby not collected by our framework which collects only positive concepts. Second, treatments were not identified by YTEX because they were long phrases or contain certain punctuations such as: "reviewed the major concepts in mind body bridging" and “psychologist – psychotherapy” which might not be collected because the YTEX pipeline breaks the segments into noun phrases. Another example of complications due to long phrases is "point out consequences of impulsive care". Finally, the treatment concept may not exist in the *PTSD-KB* or exists as a variation. For example, the missed phrase "prevent relapse" is a permutation of the existent "relapse prevention" within the PTSD-KB.

TABLE IV: STATISTICS OF INDIVIDUAL TREATMENT TYPES

	Pharma-cology	**Psycho-logical**	**Psycho-Social**	**Psycho-Education**	**Case manag-ment**	**Other**
Manual_ BOCat	54	100	9	33	3	4
Detected	49	89	6	27	1	4
Undetected	5	11	3	6	2	0
sensitivity	90.7%	89.00%	66.67%	81.82%	33.33%	100%

The manual annotation was performed by annotators who were qualified and with sufficient content knowledge to do the task. On the other hand, automated systems are characterized with reliability and effectiveness once the theoretical framework is correctly identified. Using our framework, we have shown this to be true. Thus, some treatment type mentions, which in fact exist in the notes, were not manually annotated, but were detected by our framework. We examined a sample of those notes and found that treatment type mentions that were automatically annotated but missed by annotators did appear in the notes. Those types belong to Pharmacotherapy, Psychological, Psycho-Social and Psycho-Education in the following counts of missed treatment mentions: 3, 21, 3, 36, respectively.

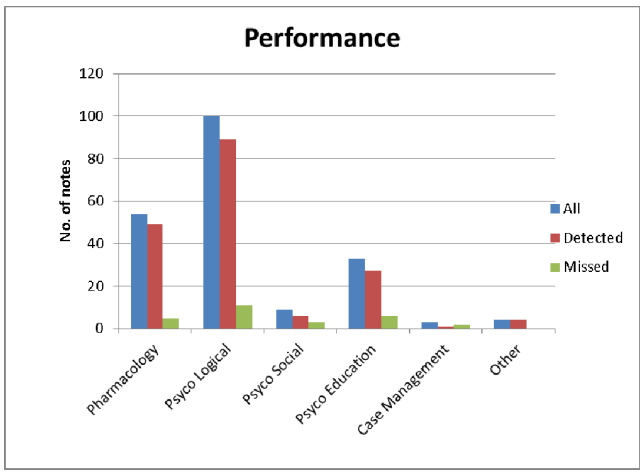

Figure 2: Distribution of Detected and Undetected treatment types

Our framework identified a potential limitation in the proficiency of any automated system that has a component of contextual knowledge, namely the *PTSD-KB*. It is crucial that *PTSD-KB* be comprehensive, have a wide coverage and include all PTSD related treatment concepts in as much as these factors have a great impact on the performance of MedCat. While experimenting with the development implementation, we iteratively ran YTEX to improve our detection, adding new treatment concepts to the *PTSD-KB*. Although our sample covers 125 sites of the VA and included many different styles of documentations, we believe that *PTSD-KB* should be subjected to frequent maintenance updates to provide continued reliability in MedCat.

VI. CONCLUSION

We have presented a framework for defining a novel compressed representation of clinical narrative data. We applied our framework to PTSD clinical notes to generate a parallel representation of the notes using PTSD treatment categories. We demonstrate the effectiveness of our method by comparing the treatment types being detected and undetected in each note to a humanly generated reference standard. We also show the capability of our method to identify treatment types that have not been tagged during manual annotation of the notes. We conclude that our framework is advantageous for enhanced quality in understanding patient care in PTSD. Furthermore, the categories represented in our conceptualization, span across the entire field of psychotherapeutic treatment, for most if not all, mental health disorders. This is to say that the five specific categories of treatment delineated in this effort, can also be used to capture the treatments documented for other mental health disorders, such as, Generalized Anxiety Disorder and Major Depressive Disorder. Promising use of this hierarchical categorization of concept-derived content also extends to other conceptual areas within the mental health domain, such as, symptom or trauma.

We also propose that this automated representation is generalizable enough to extend across medical domains where narrative content can be rolled up into a higher-level data conceptualization to more effectively evaluate different types of treatment modalities. Thus, on a larger scale, the ability to automatically aggregate detailed treatment concepts, within this and other medical domains, into a hierarchical conceptual model could have a useful impact on the evaluation of the effectiveness of different treatment types offered for other medical conditions when combined with knowledge about patient presentation or symptomatology.

ACKNOWLEDGMENT

This study was funded by Veteran Administration VA grant HIR 08-374 HSR&D: Consortium for Healthcare Informatics.

REFERENCES

[1] Aronson A. R., and Lang F.M.. "An overview of MetaMap: historical perspective and recent advances." *Journal of the American Medical Informatics Association* 2010, 17.3, 229-236.

[2] Christensen, Lee M., Harkema H., Haug P.J., Irwin J.Y., and Chapman W.. "ONYX: a system for the semantic analysis of clinical text." In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 19-27. Association for Computational Linguistics, 2009.

[3] Cully, J. A., Tolpin, L., Henderson, L., Jimenez, D., Kunik, M. E. & Peterson, L. A. 2008Psychotherapy in the veterans health administration: missed opportunities? *Psychological Services* 5 2008, (4), 320–331

[4] Cunningham H. GATE, a general architecure for text engineering. *Comput Humanit* 2004;36:223e54.

[5] Ferrucci D. A., Lally A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 200410(3-4): 327-348.

[6] Denny J. C., Randolph A. Miller A. Spickard III, Schildcrout J., Darbar D., Rosenbloom S.T., and Peterson J. F.. "Identifying UMLS concepts from ECG Impressions using KnowledgeMap." In *AMIA*

Annual Symposium Proceedings, vol. 2005, p. 196. American Medical Informatics Association, 2005.

[7] Elkin P. L., Brown S. H., Husser C. S., Bauer B. A.,Wahner-Roedler D., Rosenbloom S. T., and Speroff T. "Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists." In *Mayo Clinic Proceedings*, vol. 81, no. 6, pp. 741-748. Elsevier, 2006.

[8] Fodeh S J., Chatterjee S., Brandt C., Analysis of VA telephone call notes using topic Modeling. *American Medical Informatics Association*, 2012, 1735.

[9] Fodeh S.J., Punch W.F., Tan P.N. Combining statistics and semantics via ensemble model for document clustering. *ACM symposium on Applied Computing* 2009;1446-1450.

[10] Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association,* 11(5), 392-402.

[11] Garla, Vijay, et al. "The Yale cTAKES extensions for document classification: architecture and application." *Journal of the American Medical Informatics Association* 18.5 2011: 614-620.

[12] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18:544-51.

[13] Ogren, Philip V. "Knowtator: a protégé plug-in for annotated corpus construction." Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations. Association for Computational Linguistics, 2006.

[14] Zeng QT, Redd D., Divita G.,Jarad S.,Brandt C., Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes, *J Health Med Informat* 2011, S 3, 2.

[15] Savova G.K., Masanz J.J., Ogren P.V., et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *JAMIA* 2010;17(5):507.

[16] Seal K. H., Maguen S., Cohen B., Gima K. S., Metzler T. J. ,Bertenthal D. and Mamar, C. R. VA mental health services utilization in Iraq and Afghanistan veterans in the first year of receiving new mental health diagnoses. *Journal of Traumatic Stress*, 2010,23 (1),5–16.

[17] Fodeh S.J., Zeng Q.,Redd D.,Divita G., Brandt C.A., Clinical Note Type Analysis, *American Medical Informatics Association*, 2011, 1767.

[18] Shiner B. et al. "Automated classification of psychotherapy note text: implications for quality assessment in PTSD care." *Journal of evaluation in clinical practice* 18.3 2012,698-701.

[19] Spoont, M. R., Murdoch, M., Hodges, J. & Nugent, S. Treatment receipt by Veterans after a PTSD diagnosis in PTSD, mental health, or general medical clinics. *Psychiatric Services* 2010. 61 (1), 58–63.

[20] Stanfill M.H., Williams M., Fenton S.H., et al. S systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010;17;646-51

[21] Torii M., Kavishwar W., and Hongfang L. "Using machine learning for concept extraction on clinical documents from multiple data sources." Journal of the American Medical Informatics Association 18, no. 5 (2011): 580-587.

[22] Uzuner O, South B, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions,and relations in clinical text. J Am Med Inform Assoc 2011;18:552e6.

[23] www.nlm.nih.gov/research/**umls**/

[24] Hua X.,. Stenner P. S., Son D., Kevin B. J., Waitman L.R., and Joshua C. D.. "MedEx: a medication information extraction system for clinical narratives." *Journal of the American Medical Informatics Association* 17, no. 1 (2010): 19-24

[25] Zou Q., Wesley W. Chu,Morioka C., Leazer G.H., and Kangarloo H.. "IndexFinder: a method of extracting key concepts from clinical texts for indexing." In *AMIA Annual Symposium Proceedings*, *American Medical Informatics Association*, 2003,763.