Automated Identification of Medical Concepts and Assertions in Medical Text

Rómer Rosales, Faisal Farooq, Balaji Krishnapuram, Shipeng Yu, Glenn Fung
Knowledge Solutions, Siemens Healthcare. Malvern, PA USA

Abstract

This paper describes a machine learning, text processing approach that allows the extraction of key medical information from unstructured text in Electronic Medical Records. The approach utilizes a novel text representation that shares the simplicity of the widely used bag-of-words representation, but can also represent some form of semantic information in the text. The large dimensionality of this type of learning models is controlled by the use of a ℓ_1 regularization to favor parsimonious models. Experimental results demonstrate the accuracy of the approach in extracting medical assertions that can be associated to polarity and relevance detection.

1 Introduction

Electronic Medical Record (EMR) adoption is experiencing a rapid growth fueled by advances in Information Technology. In the United States, this is in part driven by the recent regulatory mandates and government funding. The ability to mine key, actionable information from electronic data is a key factor that motivates the increased EMR adoption from the stand point of both government and providers. Actionable information that can be regularly and systematically mined from EMRs could lead to improved operational, financial, and clinical outcomes.

However, much of the key information needed for measuring and driving process efficiencies resides in unstructured free text, and often needs to be mined and extracted into structured form. Despite the increasing emphasis on collecting key information in structured fields of EMRs, in general there will often be the need to mine other - as yet unthought of - information from free text documents. This is primarily because it is impossible to anticipate and precisely identify/define all the relevant information that would be useful for clinical, operational, and financial needs that may arise in the future. Given the constantly changing nature of medical knowledge in the form of evidence-based treatment guidelines, the definitions of key information elements also change rapidly over time. Thus, the need to focus on approaches that can easily be adapted to these changes.

While the need to automatically extract key information from medical text has been widely accepted[1], EMR systems often lack a flexible, user configurable, solution for this type of automation. Two of the reasons for this are: (1) the challenging nature of the text *understanding* task by itself and (2) the level of flexibility required (at the user end) for a solution to be practical. Open source NLP software toolkits available for extraction of information from text generally require expertise beyond that provided by medical personnel to configure new questions/concepts. At the other end of the spectrum, simple search based tools allow doctors and nurses to search for documents of patients that contain some phrases, sometimes expanding the search based on medical ontologies to get a slightly large set of documents. However such search based approaches are fundamentally limited in their ability to answer complicated clinical questions, and do not lend themselves to efficient and optimized processes for information extraction.

In this paper we describe a general machine learning approach that together with a simple graphical tool, allows medical personnel to rapidly mine their existing records to answer or setup (configure) complicated clinical questions without requiring complicated software development. Thus not only do end users obtain very high accuracy in their answers to clinical questions, they can also do it with very little effort or expenditure of time.

Significant amount of related work in the literature has focused on areas such as radiology and pathology reports[1]; for instance, automatic structuring of radiology reports[11]. More recently, researchers are making progress in the automated classification of clinical free text to coding[9] and applying machine learning and natural language processing for text mining in systems like MedIE[13] and CLEF[8].

Friedman et al.[4] discuss the potential of using NLP techniques in the medical domain, and also provides a comparative overview of the state-of-the-art NLP tools applied to biomedical text. Literature in[4,5] provide a survey of various approaches to information extraction from biomedical text including named entity tagging and extracting relationship between different entities and between different texts. Of direct relevance is the analysis of doctors' dictations by Chapman[2], which

[1]For example, the AMIA i2b2/VA text mining challenge

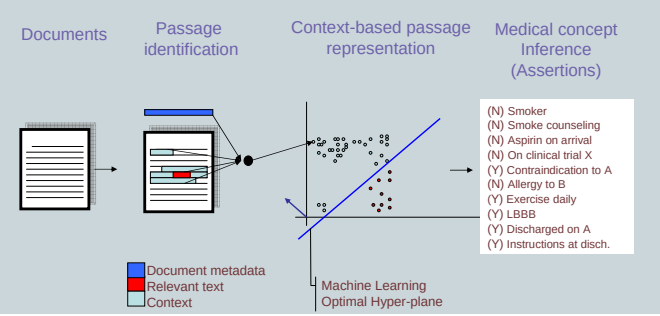

Figure 1: Identification of the basic document elements, representation of the passage in a coordinate system (along an optimal hyper-plane classifier), and various examples of inferred concepts.

identifies the seven most common uses of semantic negation in the text.

Some of the drawbacks of these works include; i) based on hard-coded rules making them hard to maintain and difficult to adapt, ii) tuned for specific tasks such as breast care reports or pathology reports; thus failing to generalize, iii) based on institution-specific styles, rules and guidelines. In all fairness, this is partially because high quality, labeled datasets of clinical documents have not been available, due to privacy laws and costs. More recently the Informatics for Integrating Biology and the Bedside (i2b2) Center (NIH funded effort) has been releasing de-identified discharge summaries for various shared challenge NLP tasks.

2 Overview of the Approach

A fundamental task in text analysis is to ascertain or infer that a piece of text (a sentence, passage, or document) refers to a particular, given topic or concept. The topic can vary widely; for example, the passage *The patient is allergic to aspirin* is not likely to refer to whether aspirin was prescribed to the patient but it is more likely to refer to whether there are any drug contraindications assessed for the patient.

A virtually equivalent task is to determine the polarity of the text; that is, whether the piece of text represents positive or negative evidence about the topic. For example, *The patient is allergic to aspirin* clearly provides positive evidence about aspirin allergies, while *The patient does not have any know allergies* provides negative evidence about drug (aspiring in this case) contraindication. A solution for this text inference problem would allow us to solve (or help solving) more complex tasks requiring text analysis. This will in turn, allow high-throughput analysis in large collection of patients. In order to achieve this, we use a text representation, learning, and inference (or classification) components illustrated in Fig. 1.

Text representation: the text to be analyzed is represented based on a combination of the document metadata (document type, date, formatting information) and contextual information. For a passage of interest, the context is defined as the section the passage is in, the distribution of words in the passage, and the relationships between these words.

Learning from text: given the large variety of medical concepts or topics of interest, the final user needs to be able to easily define and extend the system to address new concepts. We use machine learning technology to learn from user-provided examples.

Text-based inference: when a document is analyzed, two main elements are identified: (1) document metadata and (2) the actual text in the document (the content). These are represented as a vector of real numbers and processed as will be explained next. This representation is designed to be appropriate for a machine learning system to learn from examples and to produce an inference or assertion about the medical concept or topic of interest.

3 Formulation

Let us consider a medical concept/event of interest. Let y be a variable representing the incidence (or assertion) of this particular concept of interest in a text source; y is defined over the domain $\mathcal{Y}$ (*e.g.,* $\mathcal{Y} = \{\text{true,false}\}$). Let $\mathbf{x}$ represent the text source, this can be a document (*e.g.,* a discharge summary) or a text passage; $\mathbf{x}$ is defined over our input space $\mathcal{X}$, which depends on the representation utilized. For example, the input could be represented as the full text or as the commonly employed bag-of-words. In this paper we described a different representation, where we will use $\mathbf{x} = [x_1, x_2, ..., x_D]^T$, with $x_i \in \mathbb{R}$.

3.1 A Model for Text and Concepts

We define a mathematical model to associate concepts and text strings as a function $f : \mathbb{R}^D \rightarrow \mathcal{Y}$. To this end, $\hat{y} \triangleq f(\mathbf{x})$ denotes our estimate of the concept incidence. Our goal is to build f automatically from a collection of N annotated examples $\mathcal{T} \equiv \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$, so that $y_i \approx \hat{y}_i$. These examples consist of the input text and their associated annotation for the particular concept of interest. Thus, we have basically formulated our concept identification problem as a machine learning classification problem.

Text Representation Many alternatives have been suggested to represent free text in NLP tasks. The most commonly used representation is the bag-of-

words (bow) representation. Using bow, a dictionary of T terms is built and the input text is represented as a T dimensional binary vector $\mathbf{x}$, where $x_i \in \{0, 1\}$. We let $x_i = 1$ if the i-th term occurs in the input text. While this type of representations have been widely used, they are insensitive to permutations in the terms. The text *the patient stopped drinking but is still smoking heavily* would be indistinguishable from the text *the patient is still drinking heavily but stopped smoking*. Thus, it would be technically impossible to reliably determine whether the assertion *the patient is currently a smoker* is either true or false.

Many NLP techniques have been proposed to try to parse the semantics of sentences. They include shallow parsing, semantic trees, and part-of-speech tagging methods. They each have its *pros* and *cons*, but in general they require further computational resources and have limited accuracy.

In this paper we propose an alternative representation that, in addition of being simple (as demonstrated later), it can still be useful at representing the semantics/context that is critical for correctly handling common medical assertions. We start by building a dictionary of terms and phrasal terms (terms that contain a combination of words, such as *cardiovascular disease*).

Given a passage string $\mathbf{S}$ with ordered terms $(s_1, s_2, ..., s_S)$ and a dictionary $\mathbf{D} = \{t_1, t_2, ..., t_D\}$; if $s_k = t_i$ we let:

$$x_i = Z \exp\{-\frac{d(t_i, C)^2}{2\sigma^2}\}, \quad (1)$$

where $d()$ is a distance function representing how far (number of terms) the i-th dictionary term (t_i) is to a concept-specific term set, denoted by C. Specifically, for each string s_k, if s_k matches the ith term in our dictionary, then x_i is computed using Eq. 1. This is a Gaussian kernel that allows us to place more emphasis on words/terms that are nearer to the target concept.

C is a collection of terms that are related to the concept, but importantly are not required to be very accurate. For example, for the smoker concept discussed before, $C = \{smoke, cigarette, tobacco, nicotine\}$. Note that in order to properly identify all the occurrences of the concept of interest we need these terms to have a high recall, but not necessarily high precision. The scalar variable $\sigma > 0$ is set proportional to our chosen passage window size. If we need to account for more context (terms) surrounding C, then σ should be set larger (and *vice versa*).

The vector $\mathbf{x}$ is further normalized by the (Gaussian) partition function $Z = 1/\sqrt{2\pi}\sigma$ to ensure consistent scaling of $\mathbf{x}$ computed for different passages.

Learning from text One overall problem of dictionary-based representations is that they are often high-dimensional (every term is represented as a dimension). This causes problems during learning as data is normally insufficient to properly learn such a flexible model. To address this drawback we will focus on estimating sparse models.

Our general model is a hyperplane classifier, and will be estimated using the large-margin criterion for classification. In particular, we will use the Support Vector Machine formalism[12], where the goal is to estimate parameters $\mathbf{w} = [w_1, w_2, ..., w_D]^T \in \mathbb{R}^D$ that are optimal for classification. In our approach this is formulated as follows:

$$\arg\min_{\mathbf{w}} \sum_i e_i + \lambda ||\mathbf{w}||_1$$
$$\texttt{s.t.} \qquad \sum_i y_i \mathbf{w}^T \mathbf{x}_i + e_i \geq 1, \quad (2)$$

where the first term in the objective function is the standard support vector machine error (distance from margin) and the second term is a ℓ_1 regularization term that favors sparse models (weighted by $\lambda \in \mathbb{R}$). While ℓ_2 regularization is also a sensible alternative, for the same non-sparse vector $\mathbf{w}$, the ℓ_1 regularization produces a larger penalty than the ℓ_2 regularization, and thus sparser solutions are expected for ℓ_1. Also, the sample complexity (number of data points needed to *properly learn*) for ℓ_1 regularization is in general lower for the same problem[7].

The problem above can be solved using various optimization approaches. Due to space limitations we will not go into the details for solving this problem and refer the readers to our proposed approach[10].

In order to learn a concept, we need the annotated examples that represent our training set $\mathcal{T}$. Thus, a user would be able to increase its repository of concepts by simply providing some examples of the concept of interest. This may seen a time consuming process at first. We have addressed this by designing a user interface that helps the user identify and annotate examples efficiently.

Concept Inference Once learning has been achieved, inference just consists of extracting the basic pieces of information from the document, as described in Sec. 2 and applying our learned function f to the input representation.

4 Experimental Evaluation

Experiments were performed using actual EMRs from various medium/large-size hospitals[2]. We designed

[2]Names undisclosed due to privacy agreements.

our experiments to work at the passage level. A passage is a sequence of word/tokens extracted from a document. Thus, $\mathbf{x}$ represents a passage-based observation. With the help of expert medical personnel (such as expert chart abstractors), we concentrated on gathering information about various medical concepts. These concepts were chosen primarily due to their prevalence in quality reporting.

We built 19 datasets, one for each concept, with a few repeated concepts (varying in the data collection process). These passages were obtained from a set of $\sim$10 million sentences by searching, in each case, for a few keywords C related to the concept of interest and provided by the medical expert. A random subset of the matching sentences were labeled by the expert and saved using a specially tailored user interface. As expected, for all the concepts, keywords were only useful at a first level retrieval. Not surprisingly, a mixture of completely irrelevant, affirmative, or negative sentences were obtained as a first pass; *e.g.,* not all patients with passages containing the keyword *smok** are actual smokers or even have a history of smoking.

In our problem, for each chosen passage, an expert labeled it with the labels T=True and F=False to indicate the following: (1) T $\rightarrow$ the concept is *present and affirmative* in the sentence (2)F $\rightarrow$ the concept is *absent* OR it is *present but negated* in the passage.

4.1 Experimental Settings

For these experiments we used a passage size of 100 tokens, centered at the concept-specific words C. The data was used as follows: for each dataset, we first divided it into two subsets, one held out for testing only (30%) and one used for training (70%). From the training subset, a portion was assigned for actual training (75%) and another portion for cross-validation (25%). The method requires one tunable parameter, namely λ. These settings were the same for all datasets. A total of 10 folds were performed where the above subsets were always randomized.

We used 23 pre-determined *compound* dictionary entries (features) as our initial dictionary (common to all tasks described here). They are *compound* entries because in order to account for synonyms (or words playing the same semantic role in general), related words were grouped into one dictionary element. These included keywords referring to negation (*e.g., no, not, never* etc.), nouns referring to family members, different levels of severity (categorized into normal, moderate/medium, and severe), and specially formatted strings (such as references to numbers, dates, enumerations). This was the only manual step (some level of automation is possible but not addressed here).

The above dictionary entries were used for all concepts analyzed. We enhance this initial dictionary in two ways. (1) We included terms specific to the concept of interest by automatically measuring the information content of each term (in the passages) about the label value. For this we use the information theoretic mutual information (MI) as a scoring function[3]. In this manner we added a fixed number (15) of additional terms. (2) We used the statistics of all the EMR entries from a separate database to build a medical language model once using an extension of the random-walk keyword-document approach[6]. Then, for each term we identified the probability of other terms to occur *nearby*. We included a fixed number (15) of most likely to occur terms.

4.2 Quantification of results

For each concept we compared three competing approaches. (1) In order to demonstrate the advantages of the context based representation introduced, we built a model employing the bow representation and trained it using the same large-margin based approach with a sparsity prior proposed in this paper. (2) In order to test the benefits of the automated dictionary construction components, we built another model using the proposed context based representation, but without the automated (MI + statistically related terms). (3) We built a model using our full approach, which utilizes the introduced context sensitive text representation, the automated dictionary construction steps, and the sparse model learning.

We have summarized the results in Table 1. In this table, we provide both the performance using accuracy (defined as the proportion of passages correctly classified) and the AUC (or the Area Under the Received Operating Characteristic Curve) for all the approaches. For each entry in the table, the variance is also included in parenthesis. For cross-validation (selecting the best tuning parameter), we used the AUC as target performance measure.

From the results, various points can be verified. (1) Clearly, context helps; this intuitively explains why a passage representation that utilizes contextual information (based on our Gaussian kernel in Eq. 1) is superior than a bow representation. Our proposed passage representation was clearly effective at improving the accuracy for all the concepts relative to the bow representation. (2) A manually built dictionary is a valid starting point but probably not sufficient to make automated concept learning practical (based on expert judgment). However, it significantly improves over the baseline bow model; thus illustrating the potential of the approach even with by employing a simple dictionary. (3) The automated identification of effective dic-

ID	Medical concept/event	N	BOW		Proposed (manual dict.)		Proposed (automated)	
			Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
1	Currently taking amiodarone	82	0.750 (0.007)	0.688 (0.002)	0.792 (0.005)	0.711 (0.004)	**0.833 (0.007)**	**0.795 (0.016)**
2	Joint (*e.g.,* knee) revision referred	211	0.513 (0.000)	0.607 (0.010)	0.767 (0.000)	0.787 (0.000)	**0.841 (0.003)**	**0.915 (0.001)**
3	Allergy to referred antibiotic	185	0.794 (0.000)	0.745 (0.000)	0.909 (0.001)	0.971 (0.000)	**0.982 (0.000)**	**0.989 (0.000)**
4	Currently taking referred antibiotic	182	0.735 (0.006)	0.715 (0.004)	0.815 (0.001)	0.902 (0.000)	**0.907 (0.004)**	**0.977 (0.000)**
5	Contraindication to referred antibiotic	185	0.758 (0.005)	0.786 (0.003)	0.879 (0.003)	0.914 (0.002)	**0.939 (0.003)**	**0.991 (0.000)**
6	ST elevation assessed	211	0.556 (0.003)	0.543 (0.002)	0.683 (0.002)	0.719 (0.001)	**0.862 (0.001)**	**0.902 (0.001)**
7	LBBB assessed	211	0.905 (0.001)	0.523 (0.009)	0.931 (0.002)	0.925 (0.017)	**0.995 (0.000)**	**1.000 (0.000)**
8	Currently on aspirin	496	0.782 (0.000)	0.518 (0.000)	0.901 (0.000)	0.934 (0.000)	**0.950 (0.000)**	**0.960 (0.000)**
9	Documentation for not prescribing aspirin	495	0.838 (0.000)	0.435 (0.000)	**0.964 (0.000)**	**0.979 (0.000)**	0.930 (0.000)	0.959 (0.000)
10	Discharged on aspirin medication	495	0.856 (0.000)	0.832 (0.000)	0.859 (0.000)	0.871 (0.000)	**0.919 (0.001)**	**0.964 (0.000)**
11	Atrial fibrillation and flutter	278	0.795 (0.001)	0.737 (0.000)	0.880 (0.002)	0.887 (0.000)	**0.922 (0.000)**	**0.972 (0.001)**
12	Documented atherosclerosis	325	0.605 (0.000)	0.694 (0.001)	**0.938 (0.001)**	**0.960 (0.001)**	0.930 (0.000)	0.958 (0.001)
13	Assessed for rehabilitation services	195	0.776 (0.004)	0.521 (0.000)	0.575 (0.003)	0.575 (0.009)	**0.776 (0.000)**	**0.819 (0.004)**
14	Documentation for last know-well (biased)	81	0.917 (0.000)	0.614 (0.007)	0.901 (0.091)	0.755 (0.004)	0.906 (0.147)	0.779 (0.000)
15	VTE present on arrival	508	0.679 (0.001)	0.621 (0.000)	0.774 (0.000)	0.830 (0.001)	**0.822 (0.001)**	**0.898 (0.001)**
16	Smoker one year prior to hospital arrival (I)	289	0.707 (0.001)	0.512 (0.004)	0.763 (0.002)	0.772 (0.001)	**0.825 (0.001)**	**0.902 (0.001)**
17	Documentation for antibiotic allergy	650	0.543 (0.000)	0.617 (0.002)	0.878 (0.001)	0.940 (0.000)	**0.890 (0.000)**	**0.963 (0.000)**
18	Asserted/negated antibiotic allergy	308	0.875 (0.001)	0.834 (0.001)	0.938 (0.000)	0.978 (0.000)	**0.984 (0.001)**	0.979 (0.000)
19	Smoker one year prior to hospital arrival (II)	980	0.490 (0.001)	0.500 (0.000)	0.823 (0.002)	0.913 (0.001)	**0.922 (0.000)**	**0.974 (0.000)**
	Average	316	**0.730**	**0.634**	**0.846**	**0.819**	**0.902**	**0.931**

Table 1: Accuracy and area under the ROC curve (AUC) of various classification schemes for medical concepts.

tionary entries proved very valuable for all concepts. This may point to automated dictionary construction as a promising strategy to further increase accuracy and ease of use for non-NLP experts. (4) It is possible to parse and *understand* medical free text to the extent that assertions can be properly classified for polarity (positive vs. negative) and for relevance (relevant vs. non-relevant) to a good degree even for complex concepts. Quantitatively, the proposed approach clearly outperformed the comparison models.

5 Conclusion

The general goal of this work is to develop the analysis technology to enable the effective use of unstructured free-text contained in large clinical data warehouses. This includes the automated search and identification of medical concepts and events that can impact clinical decision-making and by extension, the meaningful use of clinical data. We have developed text representations and placed the problem in the context of a mathematical programming formulation. Based on our experimental results, we believe the presented approach is a step towards achieving this general goal.

References

[1] D. Aronow and K. Coltin. Information technology applications in quality assurance and quality improvement, part II. *Joint Commission Journal on Quality Improvement*, 10:465–478, 1993.

[2] W. Chapman, W. Bridewell, P. Hanbury G., Cooper, and B. Buchanan. Evaluation of negation phrases in narrative clinical reports. *Proc. American Medical Informatics Association Symp.*, pages 105–109, 2001.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, Hoboken, 1991.

[4] C. Friedman and G. Hripcsak. Natural language processing and its future in medicine: Can computers make sense out of natural language text. *Academic Medicine*, 74(8):890–895, 1999.

[5] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[6] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *ACM SIGIR Conference*, 2001.

[7] A. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Adv. in Neural Information Processing Systems*, 2004.

[8] A. Roberts, R. Gaizauskas, and M. Hepple. Extracting clinical relationships from patient narratives. *BioNLP 2008: Curr. Trends in Biomedical NLP*, 2008.

[9] N. Sager, M. Lyman, N. Nhan, and L. Tick. Automatic encoding into SNOMED III: A preliminary investigation. *Journal of the American Medical Informatics Association*, pages 230–234, 1994.

[10] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for L1 regularization: A comparative study and two new approaches. In *European Conf. on Machine Learning*, 2007.

[11] R. Taira, S. Soderland, and R. Jakobovits. Automatic structuring of radiology free text reports. *RadioGraphics*, 21:237–245, 2001.

[12] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[13] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks. Approaches to text mining for clinical medical records. *ACM Symposium on Applied Computing*, pages 235–239, 2006.