Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-661

Automatic Detection of Inconsistencies between Free Text and Coded Data in Sarcoma Discharge Letters

Ruty RINOTT[a,1], Michele TORRESANI[b], Rossella BERTULLI[b], Abigail GOLDSTEEN[a], Paolo CASALI[b], Boaz CARMELI[a], Noam SLONIM[a]
[a] *IBM Haifa Research Labs, 165 Aba Hushi st., Haifa 31905, Israel*
[b] *Fondazione IRCCS - Istituto Nazionale dei Tumori, via Venezian, 1, Milano, Italy*

Abstract. Discordance between data stored in Electronic Health Records (EHR) may have a harmful effect on patient care. Automatic identification of such situations is an important yet challenging task, especially when the discordance involves information stored in free text fields. Here we present a method to automatically detect inconsistencies between data stored in free text and related coded fields. Using EHR data we train an ensemble of classifiers to predict the value of coded fields from the free text fields. Cases in which the classifiers predict with high confidence a code different from the clinicians' choice are marked as potential inconsistencies. Experimental results over discharge letters of sarcoma patients, verified by a domain expert, demonstrate the validity of our method.

Keywords. Clinical Decision Support, NLP, Machine Learning, EHR

Introduction

A key issue in electronic health record (EHR) design is balancing between the expressive power of storing data in free text fields, versus the benefits of using coded fields, where the clinician chooses a code from a predefined list [1]. While the use of free text facilitates rapid and relatively convenient data entry, using predefined codes can enhance EHR retrieval, mining, and analysis, and may improve communication between care givers [2]. Most EHR implementations therefore rely on both methods.

Often, both free text and coded fields are available for storing a particular data type, enabling clinicians to input potentially contradictory data [3,4]. Discordance between data in the EHR may lead to confusion and mistakes in patient care [4] and may result in spurious conclusions of applications that utilize EHR data. Identifying discordances between free text and coded fields is a challenging task, and typically requires extensive work by a domain expert [3, 4].

Here we present a method to *automatically* identify inconsistencies between free text and associated coded fields in EHRs. In short, our method works by predicting the most expected code, based on the available free text data, and highlighting cases where this prediction is different from the actual code selected by the clinician. Specifically,

[1] Corresponding Author: Ruty Rinott, E-mail: rutyr@il.ibm.com. 65 Aba Hushi st., Haifa 31905, Israel

for each free text field(s) that hold information overlapping with that captured by a coded field, we train an ensemble of machine learning classifiers [5] to predict the code. Our underlying assumption is that typically the free text data and coded fields are in concordance, allowing to properly train the machine learning classifiers. We then use the obtained classifiers to predict the code based on the free text. Cases for which all classifiers predict the same code, while the clinician has selected a different code, are marked as potential inconsistencies. We report encouraging results over a real world dataset of 734 discharge letters of Sarcoma patients that supports the validity of our proposed strategy.

1. Materials & Methods

Data. We demonstrate our methodology over anonymized discharge letters of Soft Tissue Sarcoma patients treated at Fondazione IRCCS Istituto Nazionale dei Tumori (INT) between 2006 and 2011. The letters are stored using the CDA standard as defined by the Lombardy Oncology Network and contain both free text and coded fields. The data includes 734 discharge letters, termed "documents" that span 456 treatment programs. We identified 5 coded fields that capture information which is also described to some extent in free text fields, as described in Table 1. Each row in Table 1 represents a separate inconsistency detection task, handled using the methods described below.

Table 1. Coded field and free text field(s) that contain overlapping information.

Coded Field	Free text field(s)	# of instances	# of distinct words	# of distinct codes
Presentation (clinical status)	Presentation text Disease extension Clinical Summary	261	2967	2
ICDO-T (Primary anatomic site)	Disease extension Diagnostic text Oncological history	410	3792	15
ICDO-M (Morphology)	Diagnostic text	435	385	11
Treatment program (TP)	Treatment Treatment program	128	633	8
RECIST	Clinical Summary	218	1406	5

Data Representation. For each task we define the data (X) and labels (Y). X is created from free text data, using the Bag Of Words (BOW) model [6]. Specifically, the data is represented by a matrix where the rows correspond to documents and the columns correspond to words appearing in the documents. Here we use the "sign" version of the BOW model, where X_{ij}=1 if word j appears in the relevant text fields of document i, and 0 otherwise. For tasks that involve multiple free text fields, we consider in which field the word appeared. Thus, if the word j appears in k different free text fields, we represent it by k features. Tokenizing and stemming was done using Lucene Italian Analyzer [7]. The labels, Y, are defined via the codes appearing in the relevant coded field. For two of the tasks, TP and Presentation, there was not enough information in the free text to properly train a classifier that will be able to differentiate between all codes. Hence, for these tasks, we used domain knowledge to aggregate the codes into a smaller set of high-level codes, which we used as labels.

Classification Algorithms. To reduce the number of falsely reported inconsistencies we used an ensemble of classifiers for each task [5]. We selected three classifiers which use diverse classification approaches, do not require tuning of many parameters, and provide a confidence associated with the prediction: Naïve Bayes (NB) [8], K-Nearest Neighbor (KNN) [9], and Multi-class Decision Tree (MDT) [10]. We used the Matlab implementation of NB and MDT, with the posterior probability of the most probable class (MAP) as the confidence estimation. For the KNN classifier we used Cosine similarity over the TF-IDF BOW matrix [6], with k=7. Confidence was defined as the percent of neighbors agreeing with the predicted label, weighted by their similarity to the query instance. For each classifier, we also estimated the confidence that the code selected by the clinician is inconsistent (termed inconsistency confidence), calculated as one minus the probability associated with the code chosen by the clinician. The classifiers were trained using the leave-one-out scheme: for every document, we trained the classifiers using all other documents, and then used the classifier to predict the document's code, and estimate the prediction and inconsistency confidence.

Learning in the presence of noisy labels. Unlike conventional classification settings that assume that the labels of the training data are all correct, here, we assume that the data contains inconsistencies, implying that for some instance in the training set, the label (i.e., code) provided by the clinician is incorrect. These mislabeled instances may decrease the prediction accuracy of the classifiers, even if they occur in a small fraction of the data [11]. To alleviate this problem we removed from the training data instances which are highly suspected as being mislabeled, using a method reminiscent to that suggested in [11]. Namely, after the first round of training and classification, we removed from the training data all instances for which all three classifiers predicted with high confidence a code different from that selected by the clinician. The confidence threshold was chosen for each classifier so that at most 20% of the data is considered mislabeled. We then repeated the training process, using the filtered training data. These classifiers are used to predict the label and confidence for all the documents (including those removed in the first step).

2. Results

2.1. Classification results

We first estimated our classifiers performance by comparing their predictions with the clinicians' choice. Admittedly, this is not a perfect measure, since some of the observed disagreements reflect a mistake made by the clinician. Nonetheless, as we assume that inconsistencies are relatively rare in the data, this seems like an initial reasonable measure. Table 2 summarizes these results for all tasks in terms of micro-averaged recall and precision [12].

Table 2. Micro-averaged recall and precision for each task. The first column refers to results of the classifier that achieved the highest precision (mentioned in parenthesis). The last two refer to the ensemble.

Coded Field	Precision best method	Precision ensemble	Recall ensemble
Presentation	0.95 (DT)	0.98	0.77
ICDO-T	0.74 (NB)	0.93	0.54
ICDO-M	0.91 (DT)	0.96	0.73
Treatment program	0.59 (NB)	0.64	0.34
RECIST	0.60 (DT)	0.83	0.36

When evaluating the performance of the ensemble method we considered only instances for which all classifiers predicted the same label, which lead to higher precision, at the cost of reduction in recall.

2.2. Manual validation of identified inconsistencies

For each task, we marked as potentially inconsistent cases in which all classifiers agreed on a code which was different than that chosen by the clinician. These records were then examined by expert oncologists from INT and classified into one of three options: (i) true inconsistency; (ii) false inconsistency (method error); (iii) not enough information to determine. In addition, in the three tasks for which we did not aggregate the codes, the classifiers prediction can further serve as a suggestion for the correct code. For these tasks the domain experts determined if the predicted code is indeed the correct code. The results are summarized in Table 3. To further improve the precision of our method, we can mark as potentially inconsistent only cases for which the average inconsistency confidence of the classifiers is high. In theory, this requires learning a confidence threshold for each task. In practice, since our data is relatively small, learning such a threshold may result in over fitting the data. However, to demonstrate the utility of using the inconsistency confidence, we report the percent of cases correctly predicted as inconsistent when examining only the top 50% cases, ranked according to confidence. As indicated by the last column of Table 3, this improves the precision of our method for most tasks, in some cases by a factor of 2.5.

The fairly high variance in the precision of our method for different tasks is not surprising, as the complexity of the different free text fields - and their correlation with the codified field data - highly differs from task to task. An extreme example is predicting the RECIST (Response Evaluation Criteria in Solid Tumors) code from the clinical summary field. This is a very general field clinicians use to summarize different aspects of the patients hospitalization episode, and sometimes contains little or no information regarding the tumor's response, as reflected by the large fraction of cases for which there was not enough information to determine if the code is consistent.

Table 3. Manual validation of predicted inconsistencies (inc.) for each task.

Coded Field	Cases predicted as inc.	True inc.	Not enough information.	Correct prediction	Precision using top 50% of cases
Presentation	5	3	0	N/A	0.67
ICDO-T	17	5	0	3	0. 57
ICDO-M	14	6	0	6	0.86
TP	18	15	0	N/A	0.75
RECIST	16	4	7	3	0.57

Of the 14 cases predicted as inconsistent for the ICDO-M task, 6 are related to a diagnosis of Fibromyxosarcoma. Inspection of the data reveals that there are 26 cases for which the diagnosis according to the free text was Fibromyxosarcoma. For 20 of these, the ICDO-M code selected by the clinician was “Malignant fibrous histiocytoma” while for the remaining 6, the selected code was “Sarcoma, not otherwise specified (NOS)”. The ROL version used at the time the data was collected did not contain a specific code for Fibromyxosarcoma, and thus the clinicians should have chosen the general “Sarcoma, NOS” code. Since most clinicians chose the same incorrect code, our classifier learned from that mistake, and marked the cases in which “Sarcoma, NOS” was (properly) chosen as inconsistent. This demonstrates that cases in

which clinicians repeatedly make the same mistake pose a challenge to our method. However, we believe that such cases are rare, and will usually occur when there is a lacking or ill-defined code. In such cases, as this example shows, our method highlights a potential problem, although it does not identify the inconsistent cases correctly. Indeed, in the current ROL version, a new code for Fibromyxosarcoma was added.

3. Discussion

In spite of advances in development and standardization of ontologies and terminologies, the use of free text in EHR is still very common and will probably remain so in the near future. This raises the need for automatic detection of discordances in EHR, using the free text information. Here we presented an automatic method to detect inconsistencies between free text and coded fields in clinical data and demonstrated its validity over clinical discharge summaries of STS patients.

Several directions can be pursued to improve the accuracy of our method. First, given a larger EHR dataset will probably increase the precision of the examined classifiers, and will further allow considering more advanced classification schemes. Second, much progress can be made in the free text representation. Here we used the BOW model which has the advantage of simplicity, but loses much information such as negation and factuality level. Advancing beyond BOW representations certainly merits further investigation.

Here we used our method to retrospectively examine clinical records. A valuable future implementation is in online detection, drawing the clinician's attention to inconsistencies while she is filling the record. Furthermore, our method shows promising results in terms of predicting the correct code from the free text fields. This ability can be used to suggest the correct code for the clinician, facilitating and accelerating data entry process. However, this must be done with caution, and only in cases where the prediction precision is extremely high, to avoid encouraging repeating mistakes.

References

[1] Hersh WR, Hickam DH, Leone TJ. Words, concepts, or both: optimal indexing units for automated information retrieval, Proc Annu Symp Comput Appl Med Care. 1992; 644-668.
[2] Rosenberg K, Coultas D. Acceptability of Unified Medical Language System terms as substitute for natural language general medicine clinic diagnoses. Proc Annu Symp Comp App Med Car.1994; 193-7.
[3] Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository, J Am Med Inform Assoc 2000; 7(1):42-54.
[4] Singh H. Prescription errors and outcomes related to inconsistent information transmitted through computerized order entry: a prospective study. Arch Intern Med. 2009;169(10):982-9.
[5] Dietterich TG. Ensemble Methods in Machine Learning. Proc 1st Inter Workshop on MCS. 2000; 1-15.
[6] Salton G, Developments in automatic text retrieval, Science 1991; 253:974-979.
[7] Hatcher E, Gospodnetic O. Lucene in Action. Greenwich, CT, USA: Manning Publications Co, 2004.
[8] Duda RO, Hart PE , Stork DG. Pattern Classification (2nd Edition). New York: Wiley, 2001.
[9] Cover T, Hart P. Nearest neighbor pattern classification, IEEE Trans on Info Theory.1967; 13(1):21-27
[10] Breiman, L, Friedman, Olshen R, Stone. Classification and Regression Trees. Boca Raton, FL: CRC Press, 1984.

[11] Brodley CE, Friedl MA. Identifying mislabeled training data. J Artif Intell Res. 1999; 11: 131-167.
[12] Yang Y. An evaluation of statistical approaches to text categorization. J In. Ret. 1999; 1: 67–88.