The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge[1]

Contents

[1] www.computationalmedicine.org/challenge

1 Purpose and Introduction

> **Purpose: Challenge the International Natural Language Processing (NLP) research community to create and train computational intelligence algorithms that automate the assignment of ICD-9-CM codes to clinical free text.**

It is surprisingly hard for computers to handle free text as smoothly and effectively as humans do. So far, the results of the numerous efforts to achieve this have been mixed. Indeed, at times it has appeared that the complexities of free text are such as to render the effort futile. Not so; in fact, successive attempts to address the problem of converting free text into actionable knowledge have advanced the science of natural language processing and led to demand for software that simulates and complements what people are able to do.

We are sponsoring an international challenge task on the automated processing of clinical free text. Even with advances in structured vocabularies, many hospitals continue to electronically store some patient data as free text. This practice produces terabytes of information that, beyond the clinical visit, has limited utility because of its volume and accessibility. Natural language processing can potential uncover implicit structure in this data, rendering it accessible to targeted search engines as well as special purpose systems dedicated to billing, quality assurance and discovery. This challenge offers participants an opportunity to test their untested algorithms or apply existing ones. Additionally, the challenge provides full access to a carefully anonymized body of clinical data suitable for training and testing.[1] The remaining sections outline how the data were acquired, anonymized, and prepared for release, and give guidance to potential participants on how to submit results for evaluation.

"'Results of the Challenge will be announced at the IEEE Symposium on Computational Intelligence and Data Mining - April 1-5, 2007, Honolulu, Hawaii"'

Computational Medicine Center Challenge 2007 Official Website

2 Corpus Construction

Accurate clinical data are fundamental to accurate machine learning; yet, these data are notoriously laborious to acquire because of confidentiality requirements, spotty electronic availability, and insufficient recording standards. The Challenge relies on the use of such clinical data. Described below are the methods for anonymizing the data and development of a *gold standard* corpus based on expert opinion. The corpus developed for the Challenge is intended to be used to train machine learning systems dedicated to automatic billing of medical records and other related activity. Four conditions were set forth when developing the corpora:

1. data must be anonymized to meet HIPAA standards [2],
2. develop a sample that is representative of the problem that the coder faces,
3. create a sample that so that there is enough data in the well-represented classes for the automatic labeler to do well,
4. create a sample so that a proportionate representation of low-frequency distractor classes are included.

2.1 Data Collection & Anonymization

Data for the corpus were collected from the Cincinnati Children's Hospital Medical Center's Department of Radiology. Release of the data was approved by CCHMC's Institutional Review Board. Sampling of all outpatient chest x-ray and renal procedures for a one-year period was done based on a bootstrap method. [3] The data are among those most commonly used, and are designed to provide a smallish repertoire of codes that is nonetheless able to cover a substantial proportion of pediatric radiology activity.

There are both legal and ethical responsibilities when handling and publicly releasing clinical data. In brief, these responsibilities include completely expunging data that can be used to identify a patient. [4] Annonymization of these data includes three tiers: disambiguation, anonymization, and data scrubbing. [10]

Disambiguation. The language of clinicians is fundamental to patient care, but lacks the structure and clarity necessary for natural language analysis. These clinical annotations are dense with medical jargon and acronyms that often have multiple meanings. For example, in a clinical annotation the token FT can be an abbreviation for full-term, fort (as in Fort Sumter), feet or foot, field test, full-time or family therapy. Until these free text data are disambiguated, it is not possible to be certain that other anonymization tiers are accurate. To resolve the ambiguities found in the free text, a series of clinical disambiguation rules were developed using clinical experts to translate the ambiguous terms, clinical acronyms, and abbreviations. To achieve maximum disambiguation accuracy, our analysis neighborhood was sized to use a sliding window of one trigram (3 tokens on each side of the token being evaluated). [5] This approach requires a review of the three tokens before and after the token being evaluated, e.g. FT. The experts then determined the disambiguation rules, using a majority/minority approach. That is, all instances of some-term stay as some-term unless an evaluation parameter is met. If that parameter is met, then the data are changed accordingly. Figure two then provides an example of the change.

Anonymization. To assure patient privacy clinical text that is used for non-clinical reasons must be anonymized. Anonymization should not be confused with de-identification. De-identification detects patient-specific identifiers such as patient name, age, or gender and replaces them with non-specific markers such as *. Anonymization goes a step be-

yond de-identification by attempting to replace the sensitive fields with *like* values that obscure the identity of the individual. [6]

Data Scrubbing. Once the data were disambiguated and anonymized, they were reviewed for the presence of any of the 16 possible Protected Health Information (PHI) data fields. We found that the amount of PHI routinely found in unstructured free text data fields is limited. In our case these data included patient and physician name and sometimes related dates of service; all other PHI was located in other structured database fields and could be eliminated by excluding those fields from the original data query. These results enabled us to introduce a systemic bias when necessary to retain the data. This approach retains the data's contextual value, yet still meets anonymization requirements. For example, all female names were changed to Jane, all male names were changed to John, and all surnames were changed to Johnson. In the end a fully disambiguated clinical corpus with all patient and physician names being Jane or John Johnson was created. Dates were randomly shifted. If there were multiple dates in a single record, both dates where so that the distance remained the same. For example 1/1/05 and 2/1/05 could be shifted to 2/2/06 and 3/2/06.

Visual Inspection A final visual review of the data was conducted. If a specific token was perceived as potentially violating HIPAA regulations, the entire record was deleted from the dataset. This yields a corpus of 2,200 records.

2.2 ICD-9-CM Assignment

A radiology report, as any other legal document, has many components; two parts are fundamental for assigning ICD-9-CM codes: **clinical history** - provided by an ordering physician before a radiological procedure and **impression** - reported by a radiologist after the procedure. An example of these two parts is provided below:

CLINICAL HISTORY: Cough, congestion, fever.

IMPRESSION: Increased markings with subtle patchy disease right upper lobe. Atelectasis versus pneumonia.

In the case of radiology reports, ICD-9-CM codes serve as justification to have a certain procedure performed. So an insurance company might refuse reimbursement if the codes do not provide sufficient cause to perform the procedure.

An ICD-9-CM code is a 3 to 5 digit number with a decimal point after third digit. ICD-9-CM codes are organized in a hierarchy, with the highest levels of the hierarchy simply lumping some codes together by assigning consecutive numbers, e.g.:

DISEASES OF THE RESPIRATORY SYSTEM (460-519) - PARENT

– ACUTE RESPIRATORY INFECTIONS (460-466) - CHILD

Usually at a lower level of the hierarchy medical problems are divided by a location, e.g.:

ACUTE SINUSITIS 461 - PARENT

- ACUTE MAXILLARY SINUSITIS 461.0 - CHILD
- ACUTE FRONTAL SINUSITIS 461.1 - CHILD

There are official guidelines for radiology ICD-9-CM coding. These guidelines note that every disease code requires a minimum number of digits (that is, a minimum level of specificity) before costs will be reimbursed; that a definite diagnosis should always be coded when possible, that an uncertain diagnosis should never be coded and that symptoms must never be coded when a definite diagnosis is available. Particular hospitals and insurance companies typically augment these principles with more specific internal guidelines and practices for coding. For these reasons of policy, and because of natural variation in human judgments, it is not uncommon for multiple annotators to analyze the same text but to assign different codes. We wish to understand the sources of this variation, but also need to create a definite gold standard for use in the challenge. So, in addition to having data annotated by CCHMC, a portion of the data was sent to two independent companies: **COMPANY1** and **COMPANY2**.

2.2.1 Majority annotation

We now have a total of three sets of annotations, from which we wish to create a single gold-standard. We have no reason to adopt any *a priori* preference for one annotator over another, so we fall back on democratic principles, assigning a majority annotation. For each document, we have 3 sets of codes. We define the majority annotation to consist of all and only those codes which were assigned to the document by two or more of the annotators.

Issues with majority annotation:

It could be that the majority annotation will be empty, but this will be rare, because it only happens when the codes assigned by the three annotators form disjoint sets.

In the individual annotations, the coders are required to indicate a **primary code**. By convention, the primary code is listed as the first code of the record, and has an especially strong impact on the billing process. The process of forming the annotation ignores the distinction between primary and secondary codes.

The formation of the majority annotation is illustrated in the table below:

	Company X	**Company Y**	**Company Z**	**Majority**
Document 1	AB	BC	AB	AB
Document 2	BC	ABD	CDE	BCD
Document 3	EF	EF	E	EF
Document 4	ABEF	ACEF	CDEF	ACEF

The majority annotation loses some of the information provided by the individual annotators, but reduces noise and provides a definite target against which to evaluate submissions. We have conducted an analysis of agreement statistics (not further discussed here) that suggests that this method creates a coding schema that is **similar** in many ways to the original codings and is well-matched to the need for a **most probable** ground truth in a categorization challenge.

3 Evaluation

We now explain our evaluation strategy. For the moment we are making the simplistic assumption that the majority annotation is a true gold standard and a worthwhile target for emulation. This is debatable, and we discuss it below, but for the sake of definiteness we simply stipulate that submissions will be compared against the majority annotation, and that the best possible performance is to exactly replicate that annotation. Both the gold standard and the participant submissions may (and usually will) assign more than one code to each record, so this is a multi-label classification task. This is somewhat unusual in a machine learning setting.

3.1 The main ranking measure

Our first goal is to establish a simple and fair way of ranking participants. We use a macroaveraged form of the balanced F-measure. F-measure is standardly presented in the context of an attempt by the participant to correctly assign the presence or absence of a single binary feature. [7]

	Gold-Yes	**Gold-No**	**Total**
Assign-Yes	A	B	A+B
Assign-No	C	D	C+D
Total	A+C	B+D	A+B+C+D

In this table, A represents the number of true positives (i.e. occasions where the code assigned by the participant matches one found in the gold standard), B represents the number of false positives (i.e. the number of occasions that the participant assigned a code not present in the gold standard), C represents the number of false negatives (the number of occasions that the gold standard contains a code that the participant did not provide) and D would be the number of true negatives (the number of times the gold standard and the participant agreed in failing to assign a code). The balanced F-measure is the harmonic mean of precision (P) and recall (R), standardly written as:

$$F = \frac{2PR}{P+R} \text{ where } P = \frac{A}{A+B} \text{ and } R = \frac{A}{A+C}.$$

In contrast to standard classification problems in the multi-label case the sum of all table elements is not equal to the number of documents. Notice that D is not used in F-measure calculation. This is convenient when the formula is generalized to the multilabel version. The necessary change is slight. We consider each of the possible labels in turn, treating each as a binary variable and filling the corresponding table according to the gold standard and participant's predictions for that label. The resulting contingency table is a sum over all labels. The assumption that majority annotation is a true gold standard is debatable, but it is used here for the sake of definiteness and discussed further below. As a result several codes are assigned to each record, defining a multi-label classification task with large number of classes, a rather unusual machine learning problem, although not uncommon in natural language processing.

Our first goal is to establish a simple and fair way of ranking participants. We use an extension to the multi-label case of balanced F-measure. F-measure is standardly presented in the context of an attempt by the participant to correctly assign the presence or absence of a single binary feature.

	Code in Gold	**Code not in Gold**	**Total**
Code in Guess	$\sum_{Code} A$	$\sum_{Code} B$	$\sum_{Code}(A+B)$
Code not in Guess	$\sum_{Code} C$	$\sum_{Code} D$	$\sum_{Code}(C+D)$
Total	$\sum_{Code}(A+C)$	$\sum_{Code}(B+D)$	$\sum_{Code}(A+B+C+D)$

This is the **micro average** form of the contingency table, and corresponds to a decision to weight each code assignment event equally. For simplicity we do not here pursue the implications of the undoubted fact that some treatments (and hence some code assignments) have larger financial payoffs than others. To calculate F-measure, we do not need to fill cell D or the marginals that depend on it. To fill this cell, we would have to count the number of times that the gold standard and the participant agreed in **not** assigning a code, and we would have to sum this quantity over all possible codes, which is inconvenient when the repertoire of codes is large, and impossible if we are unsure what codes are available to the annotator, as may happen in a realistic setting. In addition, if the task specification is to assign codes, the rate of agreement in **not**

assigning codes is not of great interest. One may simply ignore this cell.

Concretely, we might see the following data:

	Gold Standard	**Participant 1**	**Participant 2**	**Participant 3**
Document 1	X Y	X	X Y	Y
Document 2	X	X Z	X Y	X
Document 3	Y Z	Y Z	X Z	Z
Document 4	X Y Z	X Z	X Y Z	X

In that case the contingency table for participant 1 is

	Gold-Yes	**Gold-No**	**Total**
Assign-Yes	6	1	7
Assign-No	2	NA	NA
Total	8	NA	NA

The precision for participant 1 is $6/7 = 0.86$ and the recall $6/8 = 0.75$ so the balanced F-measure is 0.80. Similarly, the table for participant 2 is

	Gold-Yes	**Gold-No**	**Total**
Assign-Yes	7	2	9
Assign-No	1	NA	NA
Total	8	NA	NA

so the precision is $7/9 = 0.778$, the recall is $7/8 = 0.88$ and the F-measure 0.82.

Participant 3 is conservative, assigning only one code per record, but doing so perfectly. The contingency table is

	Gold-Yes	**Gold-No**	**Total**
Assign-Yes	4	0	4
Assign-No	4	NA	NA
Total	8	NA	NA

For this participant precision is 100% but recall is low, with only 50% of the necessary codes assigned. The F-measure is $(2*0.5*1.0)/(1.0+0.5) = 0.66$.

3.2 A cost-sensitive accuracy measure

While F-measure is adequate as a method for ranking, there are good reasons for wanting to augment this with a cost-sensitive measure. We want an approach that allows us to manipulate the penalties for over-coding (a false positive) and under-coding (a false negative). The penalty for under-coding is simple - the hospital loses the amount of revenue that it would have earned if it had assigned the code. Our clinical informants tell us that regulations enforce an automatic over-coding penalty of three times what a hospital earned from the erroneous code, with the additional risk of a very large negative payoff due to possible prosecution for fraud. We need something that can reflect these facts. We chose to use a method that was introduced by Boutell, Shen, Luo and Brown[8]. It is a generalized version of Jaccard's similarity metric. [9] This allows a very natural representation of differential penalties for undercoding and overcoding.

Suppose that ${}_ik^{c_g}$ is a ground truth category (a set of labels) assigned to item i by a coder c_g and ${}_ik^{c_p}$ is a predicted set of labels. Futhermore, let ${}_ifn^{c_gc_p} = {}_ik^{c_g} - {}_ik^{c_p}$ be a set of false negative labels and ${}_ifp^{c_gc_p} = {}_ik^{c_p} - {}_ik^{c_g}$ be a set of false positive labels. The α-evaluation technique [8] then scores the predictions according to the following equation:

$$score({}_ik^{c_g}, {}_ik^{c_p}) = \left(1 - \frac{\beta|{}_ifn^{c_gc_p}| + \gamma|{}_ifp^{c_gc_p}|}{|{}_ik^{c_g} \cup {}_ik^{c_g}|}\right)^{\alpha} \tag{1}$$

where α is a forgiveness rate to reflect how much to forgive errors in predicting labels, β is the penalty for false negatives and γ is the penalty for false positives, with the constraint $\beta + \gamma = 2$. Since overcoding is roughly three times as bad as undercoding we use $\gamma = 1.5$,$\beta = 0.5$ and $\alpha = 1$. This measure does not represent the possibility of prosecution for fraud.

Now let us assume that $\mathbf{I}$ is a testing data set:

$$accuracy_{\mathbf{I}} = \frac{1}{|\mathbf{I}|} \sum_{i \in \mathbf{I}} score({}_ik^{c_g}, {}_ik^{c_p}) \tag{2}$$

This measure will be used to provide an additional measure of the quality of challenge results, but the final ranking and award allocation will be based on the more familiar F-measure. The cost-sensitive measure is also potentially usable to compare the codings provided by the three companies. When comparing two companies we simply treat one as ground truth and the other as if it were a challenge participant. The accuracy measure is not symmetric when $\beta \neq \gamma$ since labels that were false positives become false negatives when companies are switched.

4 The Data

We selected for the challenge a subset of a comprehensive dataset. This was created by stratified sampling, and contained 20% of the documents in each category. Since the comprehensive data set contained 2216 documents, 25 categories and 22 labels, we included in the sample only those categories to which 100 or more documents from the comprehensive data set were assigned.

Table 8: Initial distribution of radiology reports (items) in the largest multi-label categories

CATEGORY	# REPORTS	% ALL DATA	SAMPLE SIZE
786.2	2292	12.54%	458
599.0	1332	7.29%	266
593.70	1144	6.26%	229
591	908	4.97%	182
486	865	4.73%	173
780.6	653	3.57%	131
786.50	415	2.27%	83
596.54	351	1.92%	70
788.30	362	1.98%	72
599.7	343	1.88%	69
786.07	284	1.55%	57
493.90	232	1.27%	46
780.6, 786.2	214	1.17%	43
789.00	183	1.00%	37
277.00	155	0.85%	31
518.0	160	0.88%	32
V67.09	155	0.85%	31
795.5	162	0.89%	32
V71.89	154	0.84%	31
592.0	141	0.77%	28
593.70, 599.0	128	0.70%	26
753.3	120	0.66%	24
741.90	106	0.58%	21
759.89	117	0.64%	23
596.54, 741.90	103	0.56%	21

4.1 Challenge data

The challenge data was divided in to two partitions: a training set with 978 documents and a testing set with 976. 45 ICD-9-CM labels (e.g 780.6) are used in these datasets These labels form 94 distinct combinations (e.g. the combination 780.6, 786.2). We required that any combination have at least two exemplars in the data, and we split each combination between the training and the test sets So there may be labels and combinations of labels that occur only one time in the training data, but participants can be sure that no combination will occur in the test data that has not previously occurred at least once in the training data. It is up to the participants to decide how much effort should be devoted to the modeling of low-frequency categories.

Table 9: Final distribution of radiology reports (items) in the largest multi-label categories (not all categories in the training set are shown).

CATEGORY	# REPORTS	TRAINING SIZE
786.2	310	155
599.0	193	96
593.70	161	80
780.6 786.2	151	76
486	132	66
780.6	82	41
591	81	40
786.50	65	32
596.54	62	31
788.30	58	29
599.7	50	25
786.07	48	24
V13.02	35	18
795.5	32	16
591 593.89	32	16
493.90	30	15
277.00	30	15
518.0	25	12
786.07 786.2	24	12
759.89	22	11
596.54 741.90	22	11

4.2 XML data format

The data was converted to an XML format that has two top-level subdivisions: **texts** and **codes** . Both the **text** and **code** elements that have attrributes that mark the origin of the document and of the codes. In the example below, the code origins are "CMC_MAJORITY","CMC_A","CMC_B" and "CMC_C", corresponding to the majority and the three companies, and the text origins are "CCHMC_RENAL" and "CCHMC_RADIOLOGY". This data format allows a document to contain any number of codes, including zero, and applies equally to the training set, the test set and the participant submissions. The training set data will include codes from the majority annotation as well as the three component companies. Participants may use these as they see fit (they may wish to distinguish instances where there is unanimity from controversial cases, for example) but all evaluation will be with respect to the majority code.

Participants should mark the origin of their codes with some unique identifier (such as an e-mail address) using a form like *origin="participant@mailservice.com"*. The origin code that you used will be related to the email address that you use to log in to the submisision site. Another important property of **text** and **code** element is **type** which describes the nature of the content (whether a code, a text, a report, etc.) This way there is a room for creating in future data sets that have different types of codes (ICD-9-CM, CPT, ...) and textual data taken from different departments of a hospital (radiology, emergency, surgery, ...). Below is a demonstration of the format that we provide:

```
<?xml version='1.0' standalone='yes'?>
<docs>
  <doc id="123456" type="RADIOLOGY_REPORT">
    <codes>
      <code origin="CMC_MAJORITY" type="ICD-9-CM">xxx.yy</code>
      <code origin="CMC_MAJORITY" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_A" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_B" type="ICD-9-CM">zzz.xx</code>
      <code origin="CMC_C" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_A" type="ICD-9-CM">zzz.xx</code>
      <code origin="CMC_B" type="ICD-9-CM">zzz.mn</code>
      <code origin="CMC_C" type="ICD-9-CM">zzz.xx</code>
    </codes>
    <texts>
      <text origin="CCHMC_RADIOLOGY" type="IMPRESSION">...</text>
      <text origin="CCHMC_RADIOLOGY" type="CLINICAL_HISTORY">...</text>
    </texts>
  </doc>
  <doc id="123457" type="RADIOLOGY_REPORT">
    <codes>
```

```
      <code origin="CMC_MAJORITY" type="ICD-9-CM">zzz.yy</code>
    </codes>
    <texts>
      <text origin="CCHMC_RENAL" type="IMPRESSION">...</text>
    </texts>
  </doc>
  <doc id="123458" type="RADIOLOGY_REPORT">
    <codes>
      <code origin="CMC_MAJORITY" type="ICD-9-CM">zzz.yy</code>
    </codes>
    <texts>
      <text origin="CCHMC_RENAL" type="CLINICAL_HISTORY">...</text>
    </texts>
  </doc>
  <doc id="123459" type="UROLOGY_REPORT">
    <codes>
      <code origin="CMC_MAJORITY" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_A" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_B" type="ICD-9-CM">zzz.yy</code>
      <code origin="CMC_C" type="ICD-9-CM">zzz.qq</code>
    </codes>
    <texts>
      <text origin="CCHMC_RENAL" type="CLINICAL_HISTORY">...</text>
    </texts>
  </doc>
</docs>
```

and a sample of what a participant called Isaac Newton (presumed to be the only person likely to use the tag isaac.newton@royalmint.gov.uk) might have chosen to provide as a submission.

```
<?xml version='1.0' standalone='yes'?>
<docs>
  <doc id="123456" type="RADIOLOGY_REPORT">
    <codes>
      <code origin="isaac.newton@royalmint.gov.uk" type="ICD-9-CM">666.66</code>
    </codes>
    <texts>
      <text origin="CCHMC_RADIOLOGY" type="IMPRESSION">...</text>
      <text origin="CCHMC_RADIOLOGY" type="CLINICAL_HISTORY">...</text>
    </texts>
  </doc>
  <doc id="123457" type="RADIOLOGY_REPORT">
    <codes>
```

```
        <code origin="isaac.newton@royalmint.gov.uk" type="ICD-9-CM">666.67</code>
      </codes>
      <texts>
        <text origin="CCHMC_RENAL" type="IMPRESSION">...</text>
      </texts>
    </doc>
  </docs>
```

Participants are obliged to provide an explicit answer for every document in the test set. You may choose to provide no codes for a document, but every document that is in the submission must have a counterpart in the test set. Typically, participants will pass the **texts** element through unchanged, but add one or more **code** elements to any document they wish to classify. It is crucial that your preserve the **id** attribute on the document, since our software will use this (and not the embedded text) to match submissions with the gold standard. In the submission interface we will check that the submission has a corresponding **doc** element for each of the documents in the gold standard. It will not be possible to submit a file unless this is so.

Here is the RELAX-NG schema for our data in XML syntax. We will check validity against this schema as part of the submission process. It will not be possible to submit a file unless it conforms to this schema.

```
<?xml version="1.0" encoding="UTF-8"?>
<grammar ns="" xmlns="http://relaxng.org/ns/structure/1.0" datatypeLibrary="http://ww
  <start>
    <element name="docs">
      <oneOrMore>
        <element name="doc">
          <attribute name="id">
            <data type="integer"/>
          </attribute>
          <attribute name="type">
            <data type="NCName"/>
          </attribute>
          <element name="codes">
            <zeroOrMore>
              <element name="code">
                <attribute name="origin">
                  <data type="NCName"/>
                </attribute>
                <attribute name="type">
                  <data type="NCName"/>
                </attribute>
                <data type="NCName"/>
```

```
            </element>
          </zeroOrMore>
        </element>
        <element name="texts">
          <oneOrMore>
            <element name="text">
              <attribute name="origin">
                <data type="NCName"/>
              </attribute>
              <attribute name="type">
                <data type="NCName"/>
              </attribute>
              <data type="NMTOKEN"/>
            </element>
          </oneOrMore>
        </element>
      </element>
    </oneOrMore>
  </element>
</start>
</grammar>
```

and (for completeness) the same schema in RELAX-NG compact syntax (which is actually readable)

```
default namespace = ""

start =
  element docs {
    element doc {
      attribute id { xsd:integer },
      attribute type { xsd:NCName },
      element codes {
        element code {
          attribute origin { xsd:NCName },
          attribute type { xsd:NCName },
          xsd:NCName
        }*
      },
      element texts {
        element text {
          attribute origin { xsd:NCName },
          attribute type { xsd:NCName },
          xsd:NMTOKEN
```

```
      }+
    }
  }+
}
```

These schemas were created with the following Unix commands

```
trang -Ixml -Orng sample-data.xml sample-data.ng
trang -Ixml -Ornc sample-data.xml sample-data.nc
```

and the result was checked using libxml2 with

```
xmllint --noout --relaxng sample-data.ng sample-data.xml
```

The response should be

```
sample-data.xml validates
```

5 Important Dates and Results Submission

Please follow challenge web site for latest news. Here are some important dates:

1 Dec 2006 - Web site launched

22 Jan 2007 - Registration begins

1 Feb 2007 - Training data sets released

28 Feb 2007 - Registration ends

1 Mar 2007 - Challenge data set released

18 Mar 2007 - Last day to submit results. Due by midnight EST.

1 Apr 2007 - Challenge results announced

Please use the submission form on our website. The results files should be validated before submitting (using the xmllint tool for http://xmlsoft.org).

Sample usage of xmllint

```
xmllint --noout --relaxng challenge.ng your-data.xml
```

Moreover, submissions should contain exactly ONE origin. For example, if a participant logs in to the system is isaac.newton@royalmint.gov.uk then that should also be the

value of the **origin** attribute for the ICD-9-CM codes provided by that participant. Here is a set of reasons why an XML submission might be **rejected**:

1. The document does not validate against the RELAX-NG schema.
2. There is more than one origin tag for ICD-9-CM codes in the XML file
3. A document ID (the id tag on the **doc** element) is missing or duplicated.
4. Only one submission per registrant will be accepted.

We will provide an interface that gives preliminary feedback on validity of submissions, and prevents submissions with egregious formatting errors. If your submission passes this preliminary test, but causes other problems when evaluated, we will let you know as soon as we can, and work with you to fix any easily correctable problems.

The publication stream is not yet finalized. It will most likely include conference proceedings, journal publications and a potential book. All publications related to the Challenge will include the appropriate participant(s) as co-authors.

References

[1] Anyone who wishes to use the data for other non-commercial activities is free to do so, as long as it is referenced accordingly. A complete citation will be made soon after the close of the Challenge.

[2] U.S. Heath & Human Services 45 CFR Parts 160 and 164 Standards for Privacy of Individually Identifiable Health Information; Final Rule Federal Register: August 14, 2002 (Volume 67, Number 157)Page 53181-53273

[3] Walters SJ, Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36, Health and Quality of Life Outcomes, vol 2 (2004) 26.

[4] No Authors Listed, Standards for privacy of individually identifiable health information. Office of the Assistant Secretary for Planning and Evaluation, DHHS. Final rule. Fed Regist. 2000 Dec 28;65(250):82462-829.

[5] Sammuelsson C, Wiren M. Parsing Techniques. In: Dale R, Moisl H, Somers H, eds. Handbook of Natural Language Processing. New York: Marcel Deker; 2000:59-93.

[6] Cho PS, Taira RK, and Kangarloo H, "Text boundary detection of medical reports," Proceedings of American Medical Informatics Association, Annual Symposium, p. 998 (2002).

[7] Peter Jackson, Isabelle Moulinier. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. John Benjamins Publishing Co.

[8] Boutell M, Shen X, Luo J and Brown C, Multi-label Semantic Scene Classification, Technical Report 813, Department of Computer Science, University of Rochester, 2003 September.

[9] Gower JC, Legendre P, Metric and euclidean properties of dissimilarity coefficient, Journal of Classification, vol 3 (1986) 5-48.

[10] Pestian JP, Itert L, Andersen CL, Duch W. Preparing Clinical Text for Use in Biomedical Research. Journal of Database Management. 2005;17(2):1-12