Machine Learning of Text Analysis Rules for Clinical Records

Stephen Soderland, David Aronow*, David Fisher, Jonathan Aseltine, Wendy Lehnert
University of Massachusetts, Amherst MA
*Harvard Community Health Plan, Brookline, MA
{soderlan aronow dfisher aseltine lehnert}@cs.umass.edu

Automatically extracting information in clinical free text can make available an information resource that is largely untapped. This paper describes the BADGER text analysis system, which identifies concepts contained in a text based on linguistic context. A key component of BADGER is the CRYSTAL dictionary induction system that automatically learns text analysis rules from a set of training documents. Each of these rules is generalized as far as possible without producing errors, so that a minimum number of dictionary entries cover the positive training instances.

INTRODUCTION

Much of the vital information in clinical records is in narrative form. Although this information is accessible by searching the codes with which the text is associated, the meaning of the text is relatively inaccessible in automated systems. The growing interest in automated and integrated medical records has spurred intense research into machine learning for indexing, abstracting and understanding clinical text. While most of this work has focused on areas such as radiology and pathology reports,[1] some researchers are making progress in the automated classification of clinical free text to code.[2,3]

The National Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts in Amherst is working on a strategic research and development project to automate, to the extent possible, the assignment of ICD-9-CM codes to hospital discharge summaries. The current task is to automate sufficient understanding of the content of the summaries to label phrases in them that contain information concerning (1) diagnoses and (2) signs or symptoms of disease. The categories are divided into subcategories:

Diagnosis:	Sign or Symptom:
confirmed	present
ruled out	absent
suspected	presumed
pre-existing	unknown
past	history

Looking up isolated word meanings will not suffice to make these distinctions. Consider the word "fever" in sentences 1 and 2 in Figure 1. It should be classified as sign or symptom with subtype "present" in sentence 1, and with subtype "absent" in sentence 2. An automated system will need rules about linguistic contexts that distinguish absent from present symptoms, rules that are specific to the writing style of hospital discharge summaries.

The problem of classifying phrases becomes even more difficult when no single word in isolation conveys the target concept. Examples of this are the phrase "menstrual irregularities" in sentence 3 and "cigarette smoking" in sentence 4, both of which represent the concept absent sign or symptom. Sometimes the classification of a phrase is contrary to the usual meaning of a word in isolation. In sentence 5, the word "pain", which in isolation is a sign or symptom, is part of a phrase referring to a pre-existing diagnosis.

Figure 1. Sample discharge summary sentences

1) He continues to have fever and chills.
2) The patient denies fever, chills or dysuria.
3) The patient denies any menstrual irregularities.
4) The patient denies cigarette smoking.
5) PAST HISTORY: Chronic lower back pain.

Two systems developed by the Natural Language Processing Group at CIIR are being applied to this task: BADGER a sentence analyzer and CRYSTAL a dictionary induction tool. BADGER is a general-purpose tool that can be used to analyze newswire stories as easily as medical records. It relies on a semantic lexicon and a dictionary of text analysis rules which are specific to discharge summaries. CRYSTAL uses a machine learning approach to derive the rules from a training set of discharge summaries, finding features of the local linguistic context that reliably identify the target concepts. This paper introduces BADGER and CRYSTAL and presents the results of our work to date.

BADGER SENTENCE ANALYZER

The conceptual content of each relevant phrase in the text is represented by BADGER as a case frame called a concept node (CN). As BADGER analyzes each sentence in the text, it uses a set of rules called CN definitions to determine whether to create a concept node for each segment of text. In these experiments, concept nodes are only created for phrases referring to diagnoses or to signs or symptoms of disease.

Each CN definition specifies a set of syntactic and semantic constraints that must be satisfied for the definition to be applied. The syntactic constraints operate on the subject, verb phrase, direct or indirect objects, and prepositional phrases. Any of these constituents may be tested for a sequence of specific words, for specific semantic classes in the head noun of a phrase, or for specific semantic classes in the modifiers of a phrase. The verb can be further constrained with respect to active or passive voice.

The example shown in Figure 2 is a CN definition that identifies references to absent sign or symptom in the direct object. It has a set of syntactic and semantic constraints that must be met for the CN definition to apply:

a) the subject must include the word "patient" (of semantic class <Patient or Disabled Group>)
b) the verb must include the word "denies" in the active voice
c) the direct object must have the semantic class <Sign or Symptom>.

Figure 2. CN Definition for Absent Sign or Symptom in the Direct Object

CN-type: Sign or Symptom
Subtype: Absent
Extract from Direct Object
Active voice verb
Subject constraints:
words include "PATIENT"
head class <Patient or Disabled Group>
Verb constraints:
words include "DENIES"
Direct Object constraints:
head class <Sign or Symptom>

This CN definition would extract "any episodes of nausea" from the sentence "The patient denies any episodes of nausea". It would fail to apply to the sentence "Patient denies a history of asthma", since asthma is of semantic class <Disease or Syndrome>, which is not a subclass of <Sign or Symptom>. Other CN definitions would be needed to handle "She denies any episode of nausea" or "The patient has not had any episodes of nausea". Several hundred CN definitions may be required to cover most references to absent signs or symptoms.

We are using a semantic lexicon and semantic hierarchy derived from the Unified Medical Language Systems (UMLS) MetaThesaurus and Semantic Network. This gives us a semantic hierarchy with 133 semantic classes. We derived a lexicon of 95,000 terms and phrases from the MetaThesaurus that map into classes in the semantic hierarchy. A statistically-based semantic disambiguation component is added since many of these terms and phrases have multiple semantic classes.

Figure 3 shows another example of a CN definition, which identifies pre-existing diagnoses with a set of constraints that could be summarized as "... was diagnosed with recurrence of <body_part> <disease>":

a) the verb phrase must include the word "DIAGNOSED" in the passive voice
b) there must be a prepositional phrase with preposition "WITH" and including the sequence of words "recurrence of" as well as a head noun whose semantic class is a <disease_or_syndrome> and a modifying term whose class is a <body_part>

Figure 3. CN Definition for Disease Recurrence

CN-type: Diagnosis
Subtype: Pre-existing
Extract from Prep. Phrase "WITH"
Passive voice verb
Verb constraints:
words include "DIAGNOSED"
Prep. Phrase constraints:
preposition = "WITH"
words include "RECURRENCE OF"
modifier class <Body Part or Organ>
head class <Disease or Syndrome>

This CN definition applies to a sentence such as "The patient was diagnosed with a recurrence of laryngeal cancer". Since there are no constraints on the subject, the text segment is free to have any subject, including a relative pronoun or an omitted subject.

Constructing a dictionary of several hundred such CN definitions would be a laborious task, requiring someone who combines clinical knowledge with a deep understanding of the BADGER sentence analyzer. The following section shows how CRYSTAL eliminates this

manual knowledge engineering by automatically inducing CN definitions.

CRYSTAL DICTIONARY INDUCTION TOOL

CRYSTAL derives a dictionary of CN definitions from a set of training documents. Each clause that contains a phrase with relevant information serves as a motivating example for an initial CN definition that requires the exact sequence of words and the exact set of semantic classes as in the motivating example. CRYSTAL then relaxes these constraints to merge similar CN definitions. The CN definitions in CRYSTAL's final dictionary are generalized as much as possible without producing extraction errors on the training corpus.

The first step in dictionary creation is annotation of training texts by a clinician. Each phrase that contains information to be extracted is labeled with an appropriate CN type and subtype. The annotated texts are then segmented by BADGER to create a set of training instances. Each instance is a text segment, generally a simple clause, some of whose syntactic constituents may be tagged as positive instances of a particular CN type and subtype.

Figure 4 shows the initial CN definition derived from the sentence fragment "Unremarkable with the exception of mild shortness of breath and chronically swollen ankles." The domain expert has marked "shortness of breath" and "swollen ankles" with CN type sign or symptom and subtype present. BADGER's syntactic analysis has the word "unremarkable" as the subject and the complex noun phrase "the exception of mild shortness of breath and chronically swollen ankles" as the object of a single prepositional phrase.

Because the word "unremarkable" was not found in the semantic lexicon, it gets no semantic constraint. When a syntactic constituent is made up of multiple simple noun phrases ("the exception", "mild shortness_of_breath", "chronically swollen ankles") the semantic constraints become a list of classes, all of which must be satisfied.

Figure 4. An Initial CN Definition

```
CN-type: Sign or Symptom
Subtype: Present
  Extract from Prep.Phrase "WITH"
  Verb = <NULL>
  Subject constraints:
    words include  "UNREMARKABLE"
  Prep.Phrase constraints:
    preposition =  "WITH"
    words include  "THE EXCEPTION OF
MILD   SHORTNESS_OF_BREATH AND
       CHRONICALLY SWOLLEN
ANKLES"
    modifier class <Sign or Symptom>
    head class     <Sign or Symptom>, <Body
       Location or Region>
```

The initial CN definitions created from a training example are designed to operate properly on the motivating example, but are too tightly constrained to be useful on new sentences. CRYSTAL must generalize each initial CN definition to increase its coverage by relaxing some of its constraints. Figure 5 summarizes the CRYSTAL algorithm.

CRYSTAL decides which constraints are essential by finding a similar CN definition, retaining the constraints the two have in common, and dropping all other constraints. If word constraints from the two definitions have an intersecting string of words, the unified word constraint is that intersecting string. Otherwise the word constraint is dropped. Unifying two class constraints may involve moving up the semantic hierarchy to find a common ancestor of classes in the two constraints

Figure 5. The CRYSTAL Algorithm

```
Initialize Dictionary and Training Instances
Database
Do for each initial CN definition in Dictionary
   D = an initial CN definition
   Loop
      D' = the most similar CN definition to D
      U = the unification of D and D'
      Test the coverage of U in Training
Instances
      If the error rate of U > Tolerance
           exit loop
      Set D = U
   Add D to the Dictionary
Return the Dictionary
```

The resulting generalization is tested for accuracy on the set of training instances. If the generalized CN definition is valid, the process is repeated, until further relaxation would lead to a bad CN definition. A user-defined error tolerance parameter is used to decide whether a

proposed CN definition has an acceptable error rate.

This discussion has omitted some issues of how to find the most similar CN definitions or test the coverage and error rate of a proposed CN definition against the training instances efficiently. Please refer to Soderland et al. for a more detail.[4]

EXPERIMENTAL RESULTS

BADGER has been tested on a corpus of 385 hospital discharge summaries, averaging just under one thousand words each, which produced 14,719 training instances with 2,122 positive instances of "diagnosis" and 6,047 positive instances of "sign or symptom". CRYSTAL induced a dictionary of all CN types and subtypes from this training set in 10 minutes of clock time on a DEC ALPHA AXP 3000 using 45 MB of memory.

The experimental methodology was to repeatedly partition the annotated texts into a training set and a blind test set. Dictionaries for each CN type and subtype were induced from the training set and then evaluated on the test set. Performance is measured here in terms of recall and precision, where recall is the percentage of possible phrases that the dictionary extracts and precision is the percentage correct of the extracted phrases. For example if there are 5,000 phrases that could possibly be extracted from the test set, but the dictionary extracts only 3,000, recall is 60%. If the dictionary extracts 4,000 phrases, of which only 3,000 are correct, precision is 75%.

The choice of error tolerance parameter has a significant impact on performance and can be used to manipulate a tradeoff between recall and precision. Figure 5 shows performance of a dictionary of CN definitions for "sign or symptom" of any subtype, where the error tolerance is varied from 0.0 to 0.4. The results shown are the averages of 50 random partitions of the corpus into 90% training and 10% test documents for each error tolerance.

The results in Figure 5 are for generalized CN definitions, those with coverage >= 2. Precision can be boosted by raising the minimum coverage threshold. At error tolerance of 0.0 the dictionary for "sign or symptom" had precision 79 and recall 44 at minimum coverage of 2; precision 87 and recall 35 at minimum coverage of 5; and precision 91 and recall 24 at minimum coverage of 10.

Figure 5. Effect of Error Tolerance Setting on Performance

The dictionary for all CN types and subtypes, induced at error tolerance 0.2 from all 14,719 training instances, had 194 CN definitions that covered 10 or more training instances, 527 that covered from 3 to 9, and 793 with coverage of 2.

Figure 6 displays recall against the number of positive training instances for the most frequent CN type and subtypes in the corpus. Recall is over 60 and still increasing at this level of training for "diagnosis" and "sign or symptom" of any subtype. With error tolerance set at .20 and minimum coverage at 2, precision remains fairly constant regardless of training size, at about 70 for most CN types and subtypes.

Figure 6. CRYSTAL Learning Curve: Increase of Coverage with Training Size

For this graph we have set the training partitions at 10%, 30%, 50%, 70%, and 90% of the 385 annotated documents. This was done 50 times for each training size and results averaged. The number of positive training instances in a partition depends on what CN type and subtype is being learned.

The difference in coverage for "symptom, absent" and "symptom, present" is due to a limitation in negation handling by the current version of CRYSTAL. CRYSTAL can require the word "no" or the word "without" in its constraints, which allows CN definitions that identify absent symptoms. There is no mechanism to require the absence of "no" to ensure an affirmative sense for present symptoms. The next version of CRYSTAL will explicitly handle negation and allow a constraint that a phrase be in affirmative or negative mode.

Figure 7 shows representative CN definitions to give a sense of what CRYSTAL is learning.

Figure 7. Representative CRYSTAL CN Definitions

1) Extract sign or symptom present from direct object
Verb: "REVEALED"
Direct object: "A" <Finding>
2) Extract sign or symptom absent from direct object
Verb: "REVEALED"
Direct object: "NO"
3) Extract sign or symptom absent from prep. phrase
Preposition = "WITHOUT"

CN Definition 1 in Figure 7 will apply to cases where a test revealed "a mass", "a pleural effusion", or "a murmur" and correctly identify the direct object as a present sign or symptom. The requirement of the exact word "a" eliminates cases where the test revealed "no murmurs", "normal bowel sounds", and other phrases that should be labeled as absent, rather than present, sign or symptom.

CN Definition 2 includes an exact word constraint "no" to identify absent sign or symptom. This has over 90% accuracy with no semantic constraints. CN Definition 3 also has no semantic constraints and finds absent sign or symptom in a prepositional phrase with the preposition "WITHOUT". This CN definition covers 229 training instances with 13% error rate. If a constraint is added to require the class <Finding> in the prepositional phrase, coverage drops to 75 with 5% errors.

The lack of semantic constraints on most of the high coverage CN definitions may be attributed to gaps in the semantic lexicon. We have concentrated on developing CRYSTAL's mechanism for learning generalized CN definitions and have not yet addressed the issue of adapting and augmenting for our particular application the UMLS lexicon, which lacks words such as "lesions", "rate", "rhythm", "tenderness", and "distention".

CONCLUSIONS

The BADGER text analysis system is able to use extraction rules automatically derived by the CRYSTAL dictionary induction tool. When CRYSTAL is presented with a training set of hand-annotated hospital discharge summaries, it uses machine learning techniques to derive a set of extraction rules. The goal of CRYSTAL is to find the minimum set of generalized rules that cover all of the positive training instances and to test each proposed rule against the training corpus to ensure that the error rate is within a predefined tolerance.

The requirements of CRYSTAL are a sentence analyzer, a semantic lexicon that maps individual words into classes in a semantic hierarchy, and a set of annotated training texts. CRYSTAL uses the training text to build a fully functional conceptual dictionary that requires no further knowledge engineering.

Acknowledgments

This research is supported by NSF Grant No. EEC-9209623, State/Industry/University Cooperative Research on Intelligent Information Retrieval.

References

1. Aronow DB, Coltin KL. Information technology applications in quality assurance and quality improvement, Part II. *Joint Commission Journal on Quality Improvement.* 10:465-478, 1993.

2. Satomura Y, do Amaral MB. Automated diagnostic indexing by natural language processing. *Medical Informatics*. 3:149-163, 1992.

3. Sager N, Lyman M, Nhan NT, Tick LJ. Automatic encoding into SNOMED III: A preliminary investigation. *Journal of the American Medical Informatics Association*. Supp:230-234, 1994.

4. Soderland S, Fisher D, Aseltine J, Lehnert W. CRYSTAL: Inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* 1995.