OHSU/Portland VAMC Team
Participation in the 2010 i2b2/VA Challenge Tasks

Aaron M. Cohen[1], Kyle Ambert[1], Jianji Yang[3], Robert Felder[3], Richard Sproat[2], Brian Roark[2], Kristy Hollingshead[2], Kari Baker[2]

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA
[2]Department of Science & Engineering, Oregon Health & Science University, Portland, Oregon, USA
[3]Portland Veterans Administration Medical Center (PVAMC), Portland, Oregon, USA

Abstract

Automated extraction of clinical concepts and relationships could have a significant impact on the use of the electronic medical record, both for improving quality of patient care, and for increasing secondary use of clinical data in medical research. The 2010 i2b2/VA NLP challenge was organized to advance the field of automated processing of clinical text. The challenge represents a milestone in the field of clinical text processing, combining the creation of a large, well-annotated clinical text collection for research, and the conduct of a large organized set of challenge tasks comparing state-of-the art clinical text processing methods from teams around the world. We participated in all three sub-tasks comprising the 2010 i2b2/VA challenge: concept identification, assertion labeling, and relation extraction. Our submissions utilized a variety of techniques, including concept-free parsing, part-of-speech tagging, multi-way concept classification, and multi-class linear support vector machine classification.

Introduction

The automated identification of semantic concepts and relations from text can have a significant impact on the quality and efficiency of clinical care. A significant barrier to implementing such methods within the clinical setting is the lack of computer-readable clinical text. Clinical reports, such as discharge summaries, are typically stored in natural language documents, rather than in more computer-friendly structured data formats. To overcome this barrier, various natural language processing (NLP) and machine learning techniques have been developed specifically for identifying concepts and relationship in free text. However much of the work to this end has been conducted using scientific textual data differs in important ways from the grammar-free, idiosyncratic text common to clinical reports. Furthermore, due to privacy concerns, access to clinical text for research purposes has been intermittent. In order to create systems optimized for deployment in a clinical setting, it is important for the Biomedical Informatics community to evaluate extant classification approaches on standard corpora of domain-specific textual data, and, if necessary, to create entirely new automated and semi-automated approaches that are optimized for handling clinical textual data.

Challenge Task Description

The goals of the Integrating Informatics with Biology and the Bedside (i2b2)/Veteran's Affairs (VA) Shared Task in Natural Language Processing for Clinical Data were three:

1. *Concept Extraction Task*. Create a system for labeling concepts (complete noun and adjective phrases) expressed in the text of clinical records into one of four categories: medical problem, treatment, test, and none.
2. *Assertion Task*. Create a system that will correctly interpret assertion statements as being present, absent, uncertain, conditional, or not associated with the patient.
3. *Relation Extraction Task*. Create a system that will identify concept relations between medical problems (P), tests (Te), and treatments (Tr), labeling them into one of nine categories: Tr improves P (TrIP), Tr worsens P (TrWP), Tr causes P (TrCP), Tr is administered for P (TrAP), Tr is not administered because of P (TrNAP), P indicates P (PIP), Te reveals P (TeRP), Te conducted to investigate P (TeCP), or one of three "none" classes, representing negative cases of co-occurance relations for pairs of concepts (noneTr, nonePIP, and noneTe).

The challenge organizers provided extensive training data for each task. Furthermore the challenge task was organized to provide the gold standard truth for each task as it was completed. We submitted the maximum of three system runs addressing each of the three challenge tasks.

Methods

We used separate methods for each of the three tasks, incorporating data and results from the earlier tasks as input to subsequent ones.

Concept Extraction Task: We applied three different methods to the concept extraction task: two parsing-based methods, and a hybrid parsing/semantic lookup method. Concept extraction systems 1 and 2 required the use of a context-free parser and a multi-class concept classifier, with system 2 also including a re-ranker after the parser and classifier. System 3, the hybrid system, used context-free parsing and lexical resources and Metamap[1] to determine concept types.

Context-Free Parsing. The concept extraction guidelines provided by the i2b2/VA challenge required concepts to fall within a noun-phrase (NP) or adjective-phrase (AP), therefore we used a statistical context-free parser to identify candidate concepts in NPs and APs. The well-known Charniak parser[2] uses a statistical model trained on a parsed treebank corpus (e.g., the Penn Treebank[3]) to provide hierarchical syntactic parses for input raw text.

We used the Wall St. Journal Treebank[3] to train the baseline model for parsing, but its use for this task required some initial text-normalization and model domain adaptation to yield parses of reasonable utility. To assist in this, we manually annotated 57 sentences from the i2b2 training corpus with full syntactic parse information enabling us to identify key areas of mismatch between the domains. To yield better parses from the Charniak parser, we constrained the parser in several key ways, making use of modifications to the parser code that enables such modification[4]:

- Part-of-speech tags were pre-assigned to approximately 160 abbreviations and acronyms found within the i2b2 corpus. For example, “po”, a common abbreviation for “by mouth” (per os), was pre-tagged as an adverb.
- The Penn WSJ Treebank was changed so that certain determiners falling outside of concepts in the i2b2 corpus (e.g., “no” and “any”) would also fall outside of base noun phrases in the original treebank. This yielded better agreement between syntactic constituents and reference concepts.
- For the i2b2 training data, the Charniak parser was constrained to require labeled concepts to be constituents in the tree—the parser was required to return a parse having at least one constituent in the tree covering the span of the labeled concept.

Once we had constrained the parser with these methods, we re-trained our parsing model on the parses resulting from the i2b2 training data (combined with the original training data, using standard adaptation techniques) to yield a parsing model that did not require test-time constraints. This increased the recall of concepts extracted from parse constituents from 0.765 to 0.962 on the training data. Overall parsing accuracy on our small 57 sentence hand-labeled sample improved from 0.467 F-measure to 0.581 using these methods.

Thus, for concept extraction systems, we used an unconstrained, domain-adapted parser to extract candidate concepts and features for use within the various classifiers. Since we were given true concept spans for those systems, for assertion and relation classifiers we again constrained the parser to find constituents agreeing with the concept spans provided. These parses allowed us to construct dependency trees that were then used as input features to our classification systems in the assertion and relation tasks, described below.

Concept classification. A noun phrase extracted from the parser could correspond to any of the three concept types – PROBLEM, TEST, and TREATMENT, or as NONE, indicating the absence of these concept types. We therefore built a four-way perceptron classifier using the SNoW learning architecture.[5] For each NP to be classified, features included the previous two words, the words in the NP itself, the following two words, the category of the node dominating the NP, and a variety of features derived from these (e.g., the presence of digits, and the presence of n-grams found in a manually-constructed table of procedures, disorders or chemical tests). Preliminary results involving cross validation on the training data suggested this system would perform very well: precision was measured at 0.80 and recall at 0.65. We also used a re-ranker on the SNoW system’s output, which incorporated concept labels from the SNoW system into the syntactic parses from the Charniak parser. We used re-ranker features as defined for syntactic parse reranking in the Charniak and Johnson re-ranker [5], slightly modified to allow for new non-terminal labels resulting from incorporating the concept labels. Preliminary results using this approach yielded an improvement over the SNoW system alone.

Hybrid Concept Extraction. The third system we submitted for the Concept Extraction task used a hybrid approach, drawing on syntactic rules, semantic type recognition resource, and unsupervised learning from the training data.

The system has a pipeline architecture, taking POS-tagged documents as input, and returning concept-labeled NPs. The pipeline consisted of five stages: section identification, key noun identification, concept identification, concept type mapping and reassignment, and NP construction and output generation.

Section identification. POS-tagged documents created using the above parsing methods, and were processed line by line, meaning that no cross-line references were allowed. First, we identified the section headings, such as ‘Discharge Medications’, ‘Examinations and Results’, ‘Past Medical History’. We took advantage of the discharge (DC) notes’ semi-structured format, and used the section headings to enhance the concept type reassignment process.

The section keyword dictionary was manually constructed by reviewing the DC summary notes and extracting specific key terms. Section categories were then identified using the keywords and certain surface patterns at the ends of sentences (e.g ':').

Key noun identification. The system next searched for key nouns, starting from the ends of sentences by identifying all 'NN' and 'NNS' POS labels.

Concept identification. The next step was to determine the semantic types of the key nouns identified above. The system mapped the key nouns to a dictionary of abbreviations built from the publically-available VA abbreviation file and terms added manually based on review of the training data. Additional knowledge bases to map the key noun included lists of medications, lab tests, and procedures, also obtained from VA resource files. If the system was unable to identify the semantic type for a key noun using these resources, it was then submitted to MetaMap, which returned possible UMLS semantic types. Only the highest-ranked semantic type was used to label a key noun.

Concept type mapping and reassignment. We also constructed a dictionary to map semantic types associated with a key noun to one of the three concept types of interest in the task. For example, *Disease or Syndrome* from UMLS was mapped to *problem*, while *medication*, from the VA list, was mapped to *treatment*. Similarly, *Diagnostic Procedure*, from UMLS was mapped to *test*. Based on this mapping, the NP identified was assigned its concept type.

Importantly, the grouping of the semantic types to the concept types was not mutually exclusive. For example, potassium, an *Inorganic Chemical*, can be either a treatment or a test. In order to disambiguate such cases, we used the previously-identified section labels, as well as the semantic types to reassign concept type based on the high likelihood that, for example, *test* concepts occur in the *Examinations and Results* section, and *treatment* concepts occur in *Discharge Medications* section.

Because the annotation guidelines considered abnormal test results as being a *problem*, rather than a 'test', we included an additional classification step. For each concept , initially labeled as *test*, the system looked for terms indicating abnormality, and reassigned these to type *problem* if any were found. The abnormality terms were manually created by review of the training data set.

Noun-phrase construction and output generation. Based on the POS tags and relation indices, noun phrases (NP) were constructed and labeled as the identified concept type in the generated output.

Assertion Labeling Task: We approached the assertion labeling task as a straightforward 6-way classification task. Every problem-concept identified in the concept extraction task was labeled with one of the six assertion labels: *present*, *absent*, *possible*, *conditional*, *hypothetical*, or *associated_with_someone_else*. While we tried a number of different classification algorithms, kernels, and approaches to multi-way classification on the data, cross-validation results on the training data showed that no approach performed better than using the libsvm[6] linear kernel with the built-in one-against-one multi-way classification wrapper. The ECOC method[7-9], used in prior i2b2 challenges, did not perform as well in cross-validation, nor did polynomial kernels, or DAG-based orderings[10] of the constituent 2-way classifiers.

We did include in our test runs a number of different feature types: *text features*, *metric features*, and *dependency features*. Text features included features based on the text of the problem concept under consideration, as well as a variable number of text tokens preceding and following it. Metric features included the normalized count of the different types of concepts in the sentence (problem, test, and treatment) as well as a total count. Dependency features were derived from our parsing algorithms used in the concept labeling task and included features such as the concept head and tail words, and the concept root word. No feature selection was performed. We submitted three systems for the assertion task, Assertion labeling *system 1* trained on text features, only, *system 2* trained on text and metric features, and *system 3* trained text, metric, and dependency features.

Relation Extraction Task: We also treated the relationship extraction task as a supervised machine learning classification task. However, this task was more complex, in that our approach needed to generate a list of potential relations, given the concept co-occurrences at each line of the input files. The fact that not each of the co-occurrences represented a true relation between a pair of concepts of the given types further increased the complexity. Therefore, the relation extraction task was treated as a multiple classification problem involving 11 possible classes: *TrIP*, *TrWP*, *TrCP*, *TrAP*, *TrNAP*, *PIP*, *TeRP*, *TeCP*, *noneTr*, *nonePIP*, *noneTe*. The "none" classes represented negative instances of relations between co-occurring pairs of concepts. For example a *noneTr* sample was a sentence co-occurring pair of problem and treatment concepts that were not part of a relation. We compared cross-validation performance using the three separate none classes against that of a system combining all none samples into a single none class, and found a slight improvement using the three separate classes (data not shown).

Feature types fell into the same three categories as in the assertion task, however, in comparison, the relation task used a much richer set of feature types. Text features included tokens in each of the included concepts, as well as features derived from tokens preceding, in between, and following the concept pair. We also included features designating the concept pair types as well as assertion types on these concepts. Dependency features included the concept head and tail words, the concept root word, part-of-speech information, and the max/min distance in the dependency parse from each of the concepts to the sentence root (usually the main verb). Metric features included a number of frequency and distance normalized features, such as the number of tokens between the concept pair, the number of concepts of different types between the pair and in the sentence, the difference in distance to the sentence root between the two concepts, as well as the distance between concepts in the pair in the dependency parse tree. No feature selection was performed.

As with our approach to the assertion labeling task, we used cross-validation to compare a number of different machine learning and wrapper methods for applying several SVMs to multi-class problems. Again, the one-against-one method built into libsvm performed the best. However, we also found that down-sampling the three none classes led to slightly improved performance. In our submitted system, we randomly sampled the none classes at a rate of 0.65 prior to training. We repeated both the sampling and training eight times, summing the confidence predictions of each trained model together to create our final class predictions.

We submitted three systems for the relation task. Relation extraction system 1 used text features only. System 2 used text features plus dependency features. System 3 used text features, dependency features, and metric features.

Results and Discussion

Since performance results from other teams are unavailable prior to the conference, we are unable to evaluate our systems against other approaches. However, in some cases, we can compare our results against those expected in light of our training data cross-validation results. Performance for all of our submitted systems is shown in Tables 1-3. The test results were obtained using the gold standard and scoring programs provided by the challenge task organizers. Cross-validation results were computed using our own software written to follow the official scoring as closely as possible.

Unfortunately, for the submitted versions of the concept extraction systems 1 and 2, the results during the final test runs were below 0.10 F-measure. Both the concept boundary identification, as well as the concept classification labeling, were significantly less accurate than we were anticipating, based on our cross-validation results. We are currently investigating the cause of this gross system failure.

System 3, our hybrid concept extraction entry, performed moderately well, achieving an F1 of 0.51. However, since it used lexical resources constructed from the training data, we were unable to perform cross-validation and therefore have nothing to compare this score with.

Each of our assertion classification systems achieved similar performances on the test data, which approximated our cross-validation results. Advanced parsing and metric-based features did not improve performance on this task.

Including dependency parse-based features led to some improvement on the relation extraction task (an approximate increase in F1 of 0.01 All three relation extraction systems performed below our cross-validation-informed expectations. Although it is not immediately clear why this was the case. Possibly, this could be due to either overtraining stemming from the large number of features in our models, or because of significant distributional differences between the training and test data sets.

Conclusion

We submitted three systems for each of the three tasks in the i2b2/VA 2010 challenge. We look forward to comparing our approaches and results with the other participants.

References

1. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
2. Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005); 2005: Association for Computational Linguistics.
3. Marcus M, Marcinkiewicz M, Santorini B. Building a large annotated corpus of English: The Penn Treebank. Comput Linguist. 1993;19(2):330.
4. Hollingshead K. Formalizing the use and characteristics of constraints in pipeline systems. Portland, Oregon: Oregon Health & Science University; 2010.
5. Carlson A, Cumby C, Rosen JL, Roth D. The SNoW learning architecture, Technical Report UIUCDCS-R-99-2101: UIUC CS Deptartment1999.
6. Chang C-C, Lin C-J. LIBSVM : a library for support vector machines. 2001 [cited 2006 March 20, 2006]; Available from: Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

7. Cohen AM. Five-way Smoking Status Classification using Text Hot-spot Identification and Error-Correcting Output Codes. J Am Med Inform Assoc. 2008 Jan/Feb 2008;15(1):32-5.

8. Dietterich TG, Bakiri G. Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research. 1995:263-86.

9. Ambert KH, Cohen AM. A System for Classifying Disease Co-morbidity Status from Medical Discharge Summaries Using Automated Hotspot and Negated Concept Detection. J Am Med Inform Assoc. 2009 Apr 23.

10. Platt J, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems. 2000;12(3):547-53.

Table 1. F1 concept and class identification test scores for the concept extraction task.

Concept Extraction Task		
System	F1 Concept Exact Span	F1 Class Exact Span
System 1	0.052	0.018
System 2	0.070	0.043
System 3	0.538	0.513

Table 2. F1 test and training cross-validation scores for the assertion labeling task.

Assertion Labeling Task					
System	Text Features	Dependency Features	Metric Features	F1 Class Exact Span	Micro-F1 Training Crossval
System 1	✓	✕	✕	0.927	0.930
System 2	✓	✕	✓	0.928	0.930
System 3	✓	✓	✓	0.926	0.929

Table 3. F1 test and training cross-validation scores for the relation extraction task.

Relation Extraction Task					
System	Text Features	Dependency Features	Metric Features	F1 Class Exact Span	Micro-F1 Training Crossval
System 1	✓	✕	✕	0.641	0.687
System 2	✓	✓	✕	0.654	0.698
System 3	✓	✓	✓	0.656	0.699