The Mayo/MITRE System for Discovery of Obesity and Its Comorbidities

Guergana Savova*, Cheryl Clark†, Jiaping Zheng*, K. Bretonnel Cohen†, Sean Murphy*, Ben Wellner†, David Harris†, Marcia Lazo†, John Aberdeen†, Qian Hu†, Christopher Chute*, and Lynette Hirschman†

*Biomedical Informatics Research, Mayo Clinic, Rochester, MN

†Human Language Technology Department, MITRE, Bedford, MA

Abstract

This paper describes the joint Mayo/MITRE system entries for the 2008 i2b2 community evaluation "Challenges in Natural Language Processing for Clinical Data" for the task of identifying obesity and its comorbidities from patient records. Our best systems result in macro-averaged F of 0.7377 and 0.6202 for the textual and intuitive labels respectively. The methods employed are a combination of machine learning and rule-based techniques.

1 BACKGROUND

Information extraction (IE) from clinical free text has a variety of use cases–from improving quality of care and medical error reduction to the identification of patient cohorts for study participation. The methods employed in IE are rule-based, machine learning, or hybrid—aiming to combine the benefits of domain expertise in crafting rule-based systems and the power of gaining knowledge from a multitude of sources through machine learning techniques. Meystre and colleagues [1] overview the state-of-the-art in clinical IE.

The second i2b2 task "Challenges in Natural Language Processing for Clinical Data"[1] focuses on identifying obesity and 15 types of comorbidities from patient records (see Table 1 for the full list of the 16 conditions). Document-level gold standard manual labels for the presence (Y), absence (N), questionable presence (Q) or unmentioned (U) status of each of the 16 conditions were provided by the challenge organizers for two types of judgments—textual and intuitive. Textual judgments are those that are based on explicit mentions in the text. Intuitive judgments are those that are based on inference from what is mentioned in the text. No linkage to particular textual evidence for the gold standard judgment was given.

We cast the i2b2 challenge problem as (1) a general Named Entity Recognition (NER) task for discovering medical disorder classes, and (2) a document classification task. We describe our methods and submissions below.

2 METHODS

2.1 Mayo Clinic IE System

The NLP group at the Mayo Clinic has built a large-scale, modular, real-time clinical IE system. Savova and colleagues [2] describe the details of the system and the methods used to build each component. The system is being used at the Mayo Clinic to discover important clinical facts from relatively loose clinical text, which are then stored in a structured database to enable many applications and use cases. The system is to be released in the public domain in 2008[2].

The system is built within the Unstructured Information Management Architecture (UIMA)[3], a framework which allows the development of text analysis systems by stringing together text processing components, or "annotators," into a pipeline. All annotations are stored in a XML UIMA data structure (XCAS). The Mayo Clinic IE system discovers clinical named entities (NEs), maps them to an ontology or terminology, and assigns values for the following attributes:

- Terminology/ontology concept code, which is the Concept Unique Identifier (CUI) from the Unified Medical Language System (UMLS)[4]. The system also maps to the SNOMED-CT terminology, which is part of the UMLS.
- Context, with values of *current*, *historyOf*, and *familyHistoryOf.*
- Status, with values of *confirmed*, *possible*, and *negated.*
- Related_to_patient, with values of *true* and *false* according to whether the information is about the patient.

For example, in the sentence "There are no complaints worrisome for recurrent metastatic oropharynx cancer", "metastatic oropharynx cancer" is mapped to the UMLS concept with a CUI of C1378462, has a context of *historyOf,* status of *negated,* and related_to_patient is *true.*

[1] https://www.i2b2.org/NLP/

[2] Supported partially by an IBM UIMA Innovation Award.

[3] http://incubator.apache.org/uima/

[4] http://www.nlm.nih.gov/research/umls/documentation.html

The main annotators within the system are a context-free tokenizer, context-dependent tokenizer, sentence detector, part-of-speech tagger, shallow parser, dictionary NE recognizer, machine learning NE recognizer, negation detector, and context assigner. The system has been applied to a number of retrieval use-cases and tasks.

2.2 Machine Learning Approaches

The Human Language Technology group at MITRE has applied several machine learning algorithms to a number of NLP and retrieval tasks. Two types of machine learning algorithms were used in the Obesity Challenge: a Maximum Entropy classifier and a Support Vector Machine (SVM).

Maximum entropy classifiers, also known as multinomial logistic regression models, are discriminative, probabilistic classifiers that model a distribution over a set of outcomes or classes, given a set of features derived from some observed data. In the context of the Obesity Challenge, the observed data are medical records and the outcomes consist of the possible judgments assigned to each record–Y, N, or Q for the **intuitive** task and Y, N, Q, or U for the **textual** task. We trained separate classifiers for each of the 16 co-morbidities, using the Maximum Entropy classifier found in Carafe[5]. Maximum Entropy classifiers benefit from the simplicity of a single hyper-parameter, commonly an L2 regularizer (a Gaussian prior), which we tuned to 1.0 on the development data.

A support vector machine (SVM) [3] is a method for classification by assigning a label to input vectors. An SVM locates the boundaries between data groups such that every member of each group is as far away from the boundary as possible (maximum-margin). In their most basic form, SVMs are binary linear classifiers, which means that they attempt to find a straight line, or hyperplane, boundary between two groups. If the boundary between two groups is *not* a straight line, linear classifiers will not work. However, SVMs can still be used in these non-linear applications by transforming the input data from its original feature space into one where a linear boundary does exist. The SVM algorithm then identifies the appropriate boundary in this new (often higher dimension) space.

SVMs are not restricted to binary classification but may be used in multi-class problems. A simple and effective approach to multiclass classification is called the one-against-one classifier. In this approach, a binary classifier is trained for each pair of classes. If there are k classes, then there are k(k-1)/2 classifiers. Each time a new vector must be classified, it is processed with all of the classifiers and assigned to the class with the most wins. This is the approach used by libsvm[6], which is the SVM tool that we used in the Obesity Challenge. We used libsvm to classify each medical record with either Y, N, of Q for the **intuitive** task or Y, N, Q, or U for the **textual** task, and we did this for each of 16 co-morbidities.

We used the radial basis kernel with libsvm. This kernel allows the user to set gamma (the width of the Gaussian) and cost (tradeoff between training error and margin). We tested values of .0001, .001, .01, 1, and 10 for gamma, and 1, 10 and 100 for the cost. We compared the results and selected the combination of gamma and cost which produced the best results.

2.2.1 Feature Selection

We created several features sets based on the CUIs generated by the Mayo Clinic IE system, the final document classification assigned by the Mayo IE system, and additional information extracted from the text.

Semantic features

CUIs were extracted from the output of the Mayo IE extraction system. These CUIs, together with their concept and status attributes, were converted to features.

Mayo system classification feature

The final classification assigned by the Mayo IE system was included as a feature.

Non-lexical feature extraction

A rule-based system based on regular expressions was built to identify a range of assays and physical measures, including vital signs and a variety of laboratory tests. A simple expert system was then used to convert the raw values of each assay into a discrete, binary feature. The vital signs that we identified and converted to features were:

- Blood pressure: We output a feature of HIGH SYSTOLIC BLOOD PRESSURE if the systolic BP was over 150, LOW SYSTOLIC BLOOD PRESSURE if it was below 90, and NORMAL SYSTOLIC BLOOD PRESSURE otherwise. We treated the diastolic blood pressure similarly, with the boundaries being 90 and 60. We also output a LOW BLOOD PRESSURE feature if the record contained the words *hypotension* or *hypotensive.*
- Body mass index, height, and weight: Surprisingly for an obesity-related data set, these were quite rare in the training data. We output a feature of HIGH BODY MASS INDEX if the

[5] http://sourceforge.net/projects/carafe

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm

BMI was over 30 and BODY MASS INDEX NOT HIGH if the BMI was less than or equal to 30, but the low incidence of heights and weights in the data made the calculation of BMI from heights and weights not evidently useful.

The laboratory assays that we identified and converted to features were:

- Blood glucose: We output one of four features: GLUCOSE OVER 200, GLUCOSE LESS THAN OR EQUAL TO 200, GLUCOSE 141 TO 200, GLUCOSE 121 TO 140, GLUCOSE 80 TO 120, and GLUCOSE BELOW 80 at the obvious boundary values. The LESS THAN OR EQUAL TO 200 feature was co-extracted with all of the others except GLUCOSE OVER 200.
- A1c: The fact that an A1c level was measured at all seems to be as indicative of the presence of diabetes as the actual value, so we output A1C ASSAY PRESENT whenever one was observed. Additionally, we output A1C LESS THAN 7 or A1C 7 OR HIGHER at the obvious boundary.
- Cholesterol: We output one of three features: HIGH CHOLESTEROL for values greater than or equal to 240, BORDERLINE HIGH CHOLESTEROL for values from 200 to 239, and NORMAL CHOLESTEROL for values below 200.
- HDL: We output one of three features: HDL POOR for values below 40, HDL BETTER for values from 40 to 59, and HDL BEST for values greater than or equal to 60.
- LDL: Although expert intuition suggested six categories for this feature, we smoothed it to three to deal with sparsity in the training data. We output LDL GOOD for values below 130, LDL BORDERLINE HIGH for values from 130 to 159, and LDL HIGH OR VERY HIGH for values greater than or equal to 160.
- Triglycerides: Again, we smoothed from the four features suggested by expert intuition to three features, lumping together the experts' *high* and *very high* into a single category. We output TRIGLYCERIDES GOOD for values below 150, TRIGLYCERIDES BORDERLINE HIGH for values from 150 to 199, and TRIGLYCERIDES HIGH for values of 200 and above.

Finally, we output one feature related to pulmonary function tests and another that can be measured by cardiac catheterization or with an echocardiogram:

- FEV1/FVC ratio: We output one of two values: FEV1 FVC RATIO LOW for values below 80, and FEV1 FVC RATIO NORMAL for values of 80 or higher.
- Ejection fraction: Any time that an ejection fraction was observed, we output the feature LVEF MEASURED. When a single numeric percentage value was given for the EF—often, it was not (e.g. the words *low* or *normal* might be used, or a range, such as *30 to 40%*), we also output LVEF NORMAL for values of 40 or greater and LVEF LOW for values below 40.

The same rule-based system was used to identify a set of medications used to treat high cholesterol. The system converts the names and descriptions of these medications to the following features:

- CHOLESTEROL LOWERING DRUG
- CHOLESTEROL LOWERING DRUG <drug name>

3 SUBMISSIONS

3.1 Textual judgments

Submission 1: Mayo Clinic IE system

The text is first processed through the Mayo Clinic IE system already described in section 2.1. A Document Zoner was built for the i2b2 challenge to identify section headings and insert section boundary tags specific to the i2b2 document formatting. It matches some sections that do not contain headings (such as the header) based on their content and format. It labels the section headings and text with XML tags.

By identifying the sections, the Zoner tool was able to provide a means to filter sections where NER mentions may be misleading, such as the Family History and Allergies section. Additionally, weight could be factored for mentions in specific sections, which could be leveraged to indicate relevance to post-processes.

As a post-processing step on the XCAS, we created a concept filter to retain only those concepts that are relevant to the 16 comorbidities in the obesity challenge. For that, we needed a set of the relevant CUIs. We asked a Mayo Clinic domain expert to find all UMLS CUIs for the 16 categories. That list was included in the filter. The final list has 334 CUIs.

The system creates category assignments at the document level by scanning through each relevant NE within the document and its attribute values. The system creates an instance of the Y category if the value of a relevant NE context is *current* and the status is *confirmed*.

The system assigns an N instance when the relevant NE status attribute has a *negated* value. It assigns an instance of the Q category when a relevant NE status has a value of *possible*. The algorithm for negation and uncertainty value assignment is based on [4]. An anchor word, in our case a named entity, is found, after which its surrounding context is scanned for negation or context trigger words until a stopping criterion is met. If a relevant NE concept was not found, then the document was assigned to the U category for that particular comorbidity.

In many documents, there might be several mentions of relevant NEs with potentially different value attributes (some of which might be due to system errors). For example, in one document there could be three mentions of obesity concepts, two of them *negated* and one *possible.* To be able to assign a final category at the document level, we implemented a simple weighted voting mechanism. The weights are based on the section in which the NE is located and its semantic type. In the example, the final obesity document-level assignment is determined by the higher of the weights for N and Q categories. Ties are resolved by assigning default values which are Y if there is one in the tie, and N otherwise.

Submission 2: MaxEnt with Mayo Clinic IE System features

Documents are classified by a Maximum Entropy classifier trained on Challenge textual judgment training data. Its features include (1) the CUIs for the relevant named entities identified by the Mayo Clinic IE System, (2) the final category assignments made by the Mayo Clinic System, and (3) features derived from the extracted mentions of vital signs, laboratory assays, and cholesterol-lowering medications.

Submission 3: Hybrid system

This system combines three methods: (1) Mayo rule-based classification, (2) the Maximum Entropy classification model used for submission 2, and (3) an SVM classification model using the same feature set. The final document judgments for each co-morbidity are given by the method that performed best for that co-morbidity when run on the training data. The method selected for each co-morbidity is shown in Table 1.

3.2 Intuitive judgments

Submission 1: MaxEnt

Documents are classified by a Maximum Entropy classifier that was trained on Challenge intuitive judgment training data. Its features include (1) the extracted CUIs from the Mayo IE system, and (2) features derived from extracted vital signs, laboratory assays, and cholesterol-lowering medications.

Submission 2: MaxEnt with Mayo Clinic IE System judgments

Documents are classified by a Maximum Entropy classifier trained on Challenge intuitive judgment training data. Its features include (1) the aforementioned features for Submission 1, and (2) features based on the Mayo Clinic IE system textual judgments.

Submission 3: Hybrid system

This system combines two methods: (1) the Maximum Entropy classification model used for submission 2, and (2) an SVM classification model using the same feature set. The final document judgments for each co-morbidity are given by the method that performed best for that co-morbidity when run on the training data. Table 1 shows the method selected for each co-morbidity.

Co-morbidity	**Classifier method for hybrid textual submission**	**Classifier method for hybrid intuitive submission**
Asthma	MaxEnt	MaxEnt
CAD	Mayo	SVM
CHF	MaxEnt	MaxEnt
Depression	MaxEnt	MaxEnt
Diabetes	Mayo	SVM
GERD	MaxEnt	MaxEnt
Gallstones	MaxEnt	MaxEnt
Gout	MaxEnt	SVM
Hypercholesterolemia	MaxEnt	SVM
Hypertension	MaxEnt	MaxEnt
Hypertriglyceridemia	Mayo	MaxEnt
OA	MaxEnt	MaxEnt
OSA	MaxEnt	MaxEnt
Obesity	SVM	SVM
PVD	MaxEnt	MaxEnt
Venous Insufficiency	MaxEnt	MaxEnt

Table 1. Classifier/Co-morbidity Pairing for Hybrid Textual and Intuitive Submissions

4 RESULTS AND DISCUSSION

4.1 Textual Judgments

Table 2 presents the overall results from the three textual submissions. Tables 3, 4, and 5 break down the results per category for each submission.

	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Sub1	.9667	.7077	.9667	.7627	.9667	.7314
Sub2	.9658	.7475	.9658	.6406	.9658	.6803
Sub3	.9668	.7701	.9668	.7147	.9668	.7377

Table 2. Results on Textual Judgments

The two challenging categories are Q and N. The Q category has a very low number of instances in the training and test data (39 instances for training and 17 instances for testing). For example, there are five Q training instances and only one Q test instance for GERD; therefore, if the classification of the sole test instance is incorrect, the Q F-score will be 0, with a negative impact on the final macro F. Obesity has 3 Q test instances, all of which were misclassified by our system as U. Upon error analysis, it was not apparent how the gold standard textual judgments for documents 5 and 39 were determined, as there does not seem to be any mention of obesity evidence.

A major source for the N errors is semantic negation. For example, our system assigned the incorrect U label for obesity to documents 18 and 48, whose gold standard N label is based on textual evidence such as “severely malnourished” and “appearing very weak.” On the other hand, it is debatable whether such judgments are better suited for the intuitive category, as they require some degree of inferencing. Another source of errors is missed discovery of relevant NEs, e.g. for document 8 our system failed to discover “hyperlipidemic” in “will hold off on statin since not hyperlipidemic,” resulting in an incorrect U label.

	P	R	F
Y	.9620	.9352	.9484
N	.5443	.6615	.5972
Q	.3478	.4706	.4000
U	.9766	.9835	.9801

Table 3. Per Category Results, Textual, Submission 1

	P	R	F
Y	.9507	.9411	.9459
N	.7317	.4615	.5660
Q	.3333	.1765	.2308
U	.9741	.9832	.9786

Table 4. Per Category Results, Textual, Submission 2

	P	R	F
Y	.9546	.9402	.9474
N	.7391	.5231	.6126
Q	.4118	.4118	.4118
U	.9748	.9835	.9791

Table 5. Per Category Results, Textual, Submission 3

4.2 Intuitive Judgments

As a result of a processing error, Submissions 1 and 2 were exactly the same—neither included the Mayo Clinic IE system judgments. We expect that the inclusion of Mayo IE system judgments as features would have resulted in higher F-measures for Submission 2 because those judgments draw on syntactic and semantic analysis of the text.

As can be seen from Table 6, both the micro-averaged and macro-averaged F-measures were slightly higher for the Maximum Entropy classifier than for the hybrid that combined Maximum Entropy and SVM classifiers.

It can be seen from Table 7 that the MaxEnt classifier was unable to learn to classify the Q category from the training data, and we attribute this to two factors: (1) the very small number of training instances for this category, and (2) the broad range of factors that can lead to a judgment of “questionable.” A judgment of “questionable” might derive from the presence of textual indicators (e.g., *possible*, *likely*, etc.), but it might also be based on information found in different parts of the document, and might even require consideration of additional medical knowledge, such as the relative reliability of various diagnostic procedures.

	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Sub1	.9412	.9627	.9412	.6130	.9412	.6202
Sub2	.9412	.9627	.9412	.6130	.9412	.6202
Sub3	.9404	.9604	.9404	.6139	.9404	.6198

Table 6. Results on Intuitive Judgments

	P	R	F
Y	.9505	.8578	.9018
N	.9376	.9812	.9589
Q	0	0	0

Table 7. Per Category Results, Intuitive

The MaxEnt classifier was able to learn to classify the Y and N categories, with better recall for N than for Y. An analysis of the recall errors for the Y category indicates that recall accuracy could be increased by expanding the set of medication-related features to include not only cholesterol-lowering drugs, but also medications, procedures, and medical devices associated with other co-morbidities (such as Axid and Prilosec for GERD) and assigning higher weights to such features.

Acknowledgements

We would like to thank Donna Ihrke for her help with the comorbidity-to-CUI mapping.

REFERENCES

[1] Meystre CM, Savova GK, Kipper-Schuler KC and Hurdle JE. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med 2008; 47 Suppl 1: 128-144.

[2] Savova GK, Kipper-Schuler KC, Buntrock JM and Chute CG. (2008). UIMA-based clinical information extraction system. LREC 2008: Towards enhanced interoperability for large HLT systems: UIMA for NLP.

[3] Cortes C and Vapnik V. (1995). Support-vector network. *Machine Learning*, 20:273-297.

[4] Chapman WW, Bridewell W, Hanbury P, Cooper G, Buchanan B. (2001). Evaluation of Negation Phrases in Narrative Clinical Reports. Proc AMIA, 2001.