Using CuiTools to Identify Obesity and its Co-morbidities in Discharge Summaries

Bridget T. McInnes, MS

University of Minnesota, Minneapolis, MN, USA

Abstract

This paper describes our system, CuiTools, participation in the $I2B2$ NLP Obesity Challenge. The task was to determine whether a patient is obese and whether co-morbidities that are associated with obesity exist based on the patient's discharge summery. Our system uses a supervised learning approach in which we use lexical and biomedical features into a Support Vector Machine. We use the lexical feature unigrams and the biomedical features Concept Unique Identifiers and semantic types as assigned by the MetaMap Transfer Program. Our system resulted in an increase of 9 percentage points in micro-precision over the baseline for both the textual and intuitive judgments. We also found that using both the lexical and biomedical features increased the micro-precision of the textual and intuitive judgments by 26 and 29 percentage points respectively compared to using only the lexical features. Although, there was no difference in the macro-recall or micro results.

Introduction

The Second $I2B2$ Shared-Task is a multi-class classification task to automatically identify whether a patient is obese and if they exhibit any of the 15 specified co-morbidities that often coincide with obesity based on their discharge summery. To accomplish this task, we took a supervised learning approach. We created a separate classifier for obesity and each of the 15 different co-morbidities. Each classifier assigns a discharge summary one of the four status classes for its specific co-morbidity. The results of the 16 classifiers are then combined.

Our system, CuiTools, is based on ours and others previous work in the area of word sense disambiguation (WSD). In previous supervised learning approaches to WSD, the appropriate sense of a word that had multiple senses was identified using the lexical features, unigrams[1], and the biomedical features, Concept Unique Identifiers[2] and semantic types[3], from the Unified Medical Language System (UMLS).

The UMLS is a knowledge source that contains concepts from the biomedical and clinical domain. These concepts are enumerated via Concept Unique Identifiers (CUIs). For example, the term "obstructive sleep apnea" is mapped to the CUI "C0520679: Sleep Apnea, Obstructive". The UMLS is also encoded with different semantic and syntactic structures. Some such information includes part-of-speech, related concepts, synonyms and semantic types. A semantic type is a broad subject categorization assigned to a CUI. For example, the semantic type of "C0520679: Sleep Apnea, Obstructive" is "Disease or Syndrome".

There are differences between the $I2B2$ task and supervised WSD. The goal of WSD is to assign a sense from a predetermined set of senses to a specific ambiguous word in an instance. The goal of the $I2B2$ task is to assign a status class from a predetermined set of classes to an instance. To map the $I2B2$ task to WSD in order to use CuiTools, we create a 'fake' ambiguous word that contains only a blank space and use the status class as our 'predetermined set of senses'.

In our system, we use the lexical feature, unigrams, and the biomedical features, CUIs and semantic types, that occur frequently in the discharge summary training data to create 16 classifiers to determine the obesity and co-morbidity status of each of the test discharge summaries. In the following sections, we first described our supervised system. Second, we discuss our experiments and our results. Lastly, we discuss our conclusions.

Methods

In this section, we, first, describe the majority class baseline that we use to compare the performance of our system. Second, we describe our supervised learning approach, and the lexical and biomedical features. Third, we discuss how the features were extracted. Lastly, we discuss how the results are evaluated.

Baseline: The training data consists of 730 discharge summaries from the Research Patient Data Repository of Partners HealthCare. Each of the summaries were

annotated for the existence Obesity and each of the following co-morbidities: Diabetes mellitus (DM), Hypercholesterolemia, Hypertriglyceridemia, Hypertension, Atherosclerotic CV disease (CAD), Heart failure (CHF), Peripheral vascular disease (PVD), Venous insufficiency, Osteoarthritis (OA), Obstructive sleep apnea (OSA), Asthma, GERD, Gallstones, Depression, and Gout. Out of the 730 summaries, on average 727 discharge summaries were annotated with textual annotations, and 625 with intuitive annotations for Obesity and each of its co-morbidities.

The annotation classes are "Y" if the patient has the co-morbidity, "N" if the patient does not have the co-morbidity, "Q" if there is a question whether the patient has the co-morbidity and "U" if the co-morbidity is not mentioned in the discharge summary. There exist two types of judgments provided by the annotators: $textual$ judgments in which the annotations were determined on what was in the discharge summery, and $intuitive$ judgments in which the annotations were determined based on the implicit information in the summary.

The baseline for the textual and intuitive judgments were calculated as follows. For obesity and each of its co-morbidities, the most frequent class in the training data was assigned to each of the instances in the test data. The results for the textual and intuitive baseline can be seen in Table 1 and Table 2 respectively.

Lexical Features: Lexical features have previously been used in supervised systems to classify text.[1 4 5] Lexical features are words or N-grams that frequently occur in the training data. N-grams are a sequence of N words. We use unigrams (1-grams) in our system. For example, in the phrase "She complained of shortness of breath", the unigrams are "complained", "shortness", and "breath".

Biomedical Features: We use the biomedical features CUIs and semantic types from the UMLS. The UMLS contains over 1.5 million CUIs and 135 semantic types. Each CUI is assigned one or more semantic type in order to provide a consistent categorization of all CUIs. For example, the co-morbidity $Depression$ has the semantic type "Mental or Behavioral Dysfunction", $Hypertriglyceridemia$ has the semantic type "Finding", and $Obesity$ and the remaining co-morbidities have the semantic type "Disease or Syndrome".

CUIs provide a different type of feature information than N-grams or words. CUIs allow for multi-word terms to be included as a single feature. For example, the term $chief\ complaint$ is considered to be two separate unigram features $chief$ and $complaint$ but because it maps to the CUI "C0277786: Chief Complaint" in the UMLS it would be considered a single feature. Similarly, $obstructive\ sleep\ apnea$ would be considered three separate unigram features but a single biomedical feature.

CUIs also allow for a type of normalization. For example, $heart\ failure$, $myocardial\ failure$ and $cardiac\ failure$ all map to the CUI "C0018801: Heart Failure". Therefore increasing the number of times "C0018801: Heart Failure" is seen in the training data.

Using CUIs also allows for a type of feature selection to take place although it is a question of whether this is an advantage or disadvantage. Terms that do not get mapped to a concept in the UMLS get excluded therefore we are theoretically excluding terms that are not clinical or biomedical in nature. Although because a term does not have a corresponding CUI in the UMLS does not mean that may not be a considered a good feature or even "not clinical or biomedical".

Feature Extraction and Selection: We obtain the biomedical features using the MetaMap Transfer Program (MMTx) which is a publicly available version of the concept mapping system MetaMap.[6] MMTx maps terms in biomedical text to CUIs in the UMLS. MMTx also provides the semantic type information for each mapping.

MMTx is not capable of processing the entire discharge summary due to the summary length. Therefore, we break up each of the discharge summaries based on the header titles and process the sections separately. Examples of some header titles are: DISCHARGE MEDICATIONS, COMPLICATIONS and PRINCIPAL DIAGNOSIS.

The processing of the sections separately allowed us to experiment with using only those features extracted from specific headers during the training stage of the challenge. We found that extracting features from section headers that contain the words: MEDICATION(S), DIAGNOSIS, HISTORY, EXAMINATION, and DISCHARGE obtained the highest results.

We experimented with frequency thresholds of 2, 5, 10 and 15 for both the lexical and biomedical features. We found that using a frequency threshold of 10 for all the features obtained the best results overall. A threshold of 2 and 5 was to low for our system to process all of the features and a threshold of 15 reduced the accuracy of the system.

Supervised Algorithm: We experimented with two supervised learning algorithms from the WEKA data-mining package[7] during the training stage of this challenge: i) Support Vector Machines (SMO) and ii) the

Naive Bayes algorithm. We chose these two algorithms because they have shown to perform well in other classification tasks such as word sense disambiguation. When analyzing these two algorithms, we found that the SMO consistently obtained a higher accuracy than Naive Bayes, and therefore, we only submitted and report the SMO results.

Evaluation: The submissions to the shared task were evaluated using the micro and macro-averaged precision, recall, and f-measure. Precision is the number of correct positive predictions (C) out of the total number of positive predictions (S), while recall is the number of positive predictions (C) out of the total number of positive instances (A). That is,

$$precision = \frac{|C|}{|S|} \quad recall = \frac{|C|}{|A|} \tag{1}$$

The f–measure is the harmonic mean of precision and recall:

$$f\text{–}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{2}$$

In micro-scoring, the precision and recall are computed over the the entire category. In macro-scoring, the precision and recall are computed per class and then averaged together. Micro-averaging gives an equal weight to the performance on every document favoring the performance on common categories. Macro-averaging gives an equal weight to the performance on every class, regardless of how rare or how common the class. Macro-averaging favors systems that do well on the smaller classes.

Results and Discussion

In this section, we analyze the results of our system for the textual and intuitive judgments.

Textual Results: Table 1 contains the micro and macro-precision, recall and f-measure for the textual judgment results of the majority class baseline, and our system using the lexical features (Lexical) and using both the lexical and biomedical features (Lexical+Biomedical).

The micro-scoring results show that both the Lexical and Lexical+Biomedical features obtain a 9 percentage point higher micro-precision than the baseline. They also show that there is not a difference in the overall micro-precision between the Lexical and Lexical+Biomedical features. Although, adding the biomedical features show an increase in micro-precision for Obesity and 8 out of the 15 comorbidities.

The macro-scoring results show that using Lexical features obtains a decrease of 41 percentage points in macro-precision but an 8 percentage point increase in macro-recall over the baseline. The results for the Lexical+Biomedical features obtain a decrease of 12 percentage points in macro-precision but an 8 percentage point increase in macro-recall over the baseline. They also show that using the Lexical+Biomedical features obtains a 29 percentage point increase in macro-precision over using the Lexical features but no increase or decrease in recall.

In this $I2B2$ challenge, the smaller classes are very small. Specifically, Q in the intuitive judgments, and Q and N in the textual judgments as seen in Table 3.

Table 3: **Textual and Intuitive Judgments**

Judgment	Y	N	Q	U
Textual	3,283	89	30	8,296
Intuitive	3,267	7,362	26	0

As previously stated, macro-averaging gives an equal weight to the performance on every class, regardless of how rare or how common the class. The small number of annotations though for the Q and N classes affect how the macro-precision is interpreted for our results. The calculation for macro-precision does not penalize for the classes that the system does not try to predict. For example, Table 4 shows the textual judgments for OSA and the judgments using the majority class baseline.

Table 4: **Textual Judgment of Co-morbidity OSA**

Textual Judgment	Y	N	Q	U
OSA gold standard	105	1	8	614
OSA baseline	0	0	0	728

Equations 3 and 4 respectively show the calculations of the macro-precision and recall for this example. The macro-precision for this class is very high, 96%, because three of the classes have 100% precision due to fact that there were no assignments for that class.

$$macro\text{–}p = \frac{1.0 + 1.0 + 1.0 + \frac{614}{728}}{4} = 0.96 \tag{3}$$

$$macro\text{–}r = \frac{0.0 + 0.0 + 0.0 + \frac{614}{614}}{4} = 0.25 \tag{4}$$

$$micro\text{–}p = \frac{614}{728} = 0.84 \; micro\text{–}r = \frac{614}{728} = 0.84 \tag{5}$$

Analysis of the Biomedical+Lexical and Lexical features show that the system primarily assigns the classes

Table 1: **Textual Results**

(Micro/Macro)	Baseline			Lexical			Lexical + Biomedical		
Disease	P	R	F	P	R	F	P	R	F
Obesity	0.59/0.90	0.59/0.25	0.59/0.18	0.78/0.88	0.78/0.39	0.78/0.39	0.80/0.90	0.80/0.40	0.80/0.40
Depression	0.86/0.93	0.86/0.50	0.86/0.46	0.91/0.87	0.91/0.71	0.91/0.76	0.89/0.78	0.89/0.74	0.89/0.76
Hypertriglyceridemia	0.98/0.99	0.98/0.50	0.98/0.50	0.98/0.99	0.98/0.55	0.98/0.59	0.98/0.99	0.98/0.55	0.98/0.59
Gallstones	0.82/0.94	0.82/0.33	0.82/0.30	0.85/0.84	0.85/0.43	0.85/0.45	0.85/0.84	0.85/0.45	0.85/0.47
OSA	0.86/0.95	0.86/0.33	0.86/0.31	0.92/0.90	0.92/0.51	0.92/0.54	0.92/0.90	0.92/0.54	0.92/0.55
Asthma	0.86/0.96	0.86/0.25	0.86/0.23	0.94/0.95	0.94/0.42	0.94/0.43	0.93/0.94	0.93/0.41	0.93/0.42
CAD	0.56/0.89	0.56/0.25	0.56/0.18	0.76/0.63	0.76/0.40	0.76/0.39	0.80/0.77	0.80/0.43	0.80/0.43
PVD	0.87/0.94	0.87/0.50	0.87/0.47	0.93/0.86	0.93/0.80	0.93/0.82	0.93/0.86	0.93/0.81	0.93/0.83
Gout	0.90/0.95	0.90/0.50	0.90/0.47	0.94/0.89	0.94/0.78	0.94/0.83	0.94/0.85	0.94/0.78	0.94/0.81
Diabetes	0.67/0.92	0.67/0.25	0.67/0.20	0.82/0.52	0.82/0.43	0.82/0.44	0.79/0.71	0.79/0.43	0.79/0.42
CHF	0.56/0.85	0.56/0.33	0.56/0.24	0.81/0.54	0.81/0.55	0.81/0.54	0.80/0.53	0.80/0.55	0.80/0.54
Venous Insufficiency	0.98/0.99	0.98/0.50	0.98/0.50	0.98/0.49	0.98/0.50	0.98/0.49	0.98/0.99	0.98/0.50	0.98/0.50
GERD	0.86/0.96	0.86/0.25	0.86/0.23	0.91/0.92	0.91/0.39	0.91/0.40	0.88/0.87	0.88/0.36	0.88/0.37
OA	0.83/0.91	0.83/0.50	0.83/0.45	0.83/0.70	0.83/0.68	0.83/0.69	0.84/0.72	0.84/0.67	0.84/0.69
Hypercholesterolemia	0.56/0.89	0.56/0.25	0.56/0.18	0.73/0.86	0.73/0.37	0.73/0.36	0.74/0.87	0.74/0.38	0.74/0.37
Hypertension	0.75/0.92	0.75/0.33	0.75/0.29	0.79/0.48	0.79/0.48	0.79/0.48	0.80/0.82	0.80/0.50	0.80/0.49
Overall	**0.78/0.87**	0.78/0.34	0.78/0.34	**0.87/0.46**	0.87/0.42	0.87/0.43	**0.87/0.75**	0.87/0.42	0.87/0.43

Table 2: **Intuitive Results**

(Micro/Macro)	Baseline			Lexical			Lexical + Biomedical		
Disease	P	R	F	P	R	F	P	R	F
Obesity	0.57/0.79	0.57/0.50	0.57/0.36	0.78/0.77	0.78/0.77	0.78/0.77	0.79/0.79	0.79/0.79	0.79/0.79
Depression	0.78/0.89	0.78/0.50	0.78/0.44	0.80/0.71	0.80/0.66	0.80/0.68	0.77/0.66	0.77/0.66	0.77/0.66
Hypertriglyceridemia	0.95/0.97	0.95/0.50	0.95/0.49	0.94/0.56	0.94/0.51	0.94/0.52	0.94/0.56	0.94/0.51	0.94/0.52
Gallstones	0.84/0.92	0.84/0.50	0.84/0.46	0.86/0.75	0.86/0.63	0.86/0.66	0.87/0.79	0.87/0.67	0.87/0.70
OSA	0.86/0.95	0.86/0.33	0.86/0.31	0.92/0.91	0.92/0.52	0.92/0.55	0.92/0.90	0.92/0.53	0.92/0.54
Asthma	0.86/0.93	0.86/0.50	0.86/0.46	0.93/0.91	0.93/0.80	0.93/0.84	0.92/0.87	0.92/0.76	0.92/0.80
CAD	0.59/0.86	0.59/0.33	0.59/0.25	0.83/0.88	0.83/0.55	0.83/0.55	0.83/0.88	0.83/0.55	0.83/0.55
PVD	0.86/0.95	0.86/0.33	0.86/0.31	0.91/0.89	0.91/0.52	0.91/0.53	0.92/0.91	0.92/0.53	0.92/0.55
Gout	0.88/0.94	0.88/0.50	0.88/0.47	0.94/0.91	0.94/0.78	0.94/0.83	0.95/0.92	0.95/0.82	0.95/0.86
Diabetes	0.70/0.85	0.70/0.50	0.70/0.41	0.89/0.87	0.89/0.87	0.89/0.87	0.85/0.82	0.85/0.84	0.85/0.83
CHF	0.52/0.84	0.52/0.33	0.52/0.23	0.82/0.88	0.82/0.55	0.82/0.55	0.79/0.86	0.79/0.53	0.79/0.53
Venous Insufficiency	0.93/0.97	0.93/0.50	0.93/0.48	0.92/0.63	0.92/0.57	0.92/0.59	0.92/0.64	0.92/0.57	0.92/0.59
GERD	0.78/0.93	0.78/0.33	0.78/0.29	0.83/0.83	0.83/0.48	0.83/0.49	0.79/0.79	0.79/0.44	0.79/0.45
OA	0.79/0.93	0.79/0.33	0.79/0.30	0.83/0.83	0.83/0.47	0.83/0.48	0.83/0.82	0.83/0.47	0.83/0.48
Hypercholesterolemia	0.56/0.78	0.56/0.50	0.56/0.36	0.78/0.77	0.78/0.78	0.78/0.77	0.78/0.77	0.78/0.78	0.78/0.77
Hypertension	0.80/0.90	0.80/0.50	0.80/0.45	0.80/0.69	0.80/0.67	0.80/0.68	0.80/0.67	0.80/0.64	0.80/0.65
Overall	**0.77/0.82**	0.77/0.47	0.77/0.48	**0.86/0.56**	0.86/0.55	0.86/0.56	**0.86/0.89**	0.86/0.55	0.86/0.55

Y and U to the instances. The system using Biomedical+Lexical features assigns zero Q classes and six N classes to instances in CAD, CHF and Diabetes. The system using the Lexical features assigns a Q class to an instance in Diabetes and six N classes to instances in CAD, CHF, Diabetes and Venous Insufficiency. For example, analysis of the Diabetes results show that the Biomedical+Lexical features results in a 71% macro-precision whereas the Lexical features results in a 48% macro-precision. This 23 percentage point difference is due to the 100% precision for the Q class by the Biomedical+Lexical system because it did not assign a Q at all.

Intuitive Results: Table 2 contains the micro and macro-precision, recall and f-measure for the intuitive judgment results of the majority class baseline, and our system using the lexical features (Lexical) and using both the lexical and biomedical features (Lexical+Biomedical).

The micro-scoring results show that both the Lexical and Lexical+Biomedical features obtain a 9 percentage point higher micro-precision than the baseline. They also show that there is not a difference in the overall micro-precision between the Lexical and Lexical+Biomedical features.

The macro-scoring results show that using the Lexical features obtains a decrease of 26 percentage points in macro-precision but an 8 percentage point increase in macro-recall over the baseline. The results for the Lexical+Biomedical features obtain a 7 percentage points increase in macro-precision and an 8 percentage point increase in macro-recall over the baseline. They also show that using the Lexical+Biomedical features obtains a 33 percentage point increase in macro-precision over using the Lexical features but no increase or decrease in recall. The increase in macro-precision of the Lexical+Biomedical features over the Lexical features for the intuitive results is due to the same situation that was seen in the textual results.

Conclusions

This paper described our system, CuiTools, participation in the $I2B2$ NLP Obesity Challenge. Our system used a supervised learning approach in which we use lexical and biomedical features into a Support Vector Machine. We use the lexical feature unigrams and the biomedical features CUIs and semantic types as assigned by MMTx.

Our system resulted in an increase of 9 percentage points in micro-precision over the baseline for both the textual and intuitive judgments. The results for the using both the lexical and biomedical features obtain a 7 percentage points increase in macro-precision and an 8 percentage point increase in macro-recall over the baseline for the intuitive judgments. Although, there was no difference in overall micro-precision.

Acknowledgments

The author thanks Ted Pedersen, John Carlis and Lan Aronson for their comments.

Our experiments were conducted using CuiTools v0.17, which is freely available from http://cuitools.sourceforge.net.

Address for Correspondence

Bridget McInnes, University of Minnesota Twin Cities, Department of Computer Science and Engineering, 200 Union Street SE 55455 Minneapolis (MN), USA
bthomson@cs.umn.edu

References

1. M. Joshi, T. Pedersen, and R. Maclin. A comparative study of support vectors machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of 2nd Indian International Conference on AI*, pages 3449–3468, Dec. 2005.
2. B. McInnes, T. Pedersen, and J. Carlis. Using UMLS concept unique identifiers (CUIs) for word sense disambiguation in the biomedical domain. In *Proceedings of the American Medical Informatics Association Symposium*, pages 533–537, Chicago, IL, Nov. 2007.
3. G. Leroy and T.C. Rindflesch. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Informatics*, 74(7-8):573–85, 2005.
4. T. Pedersen. Determining Smoker Status using Supervised and Unsupervised Learning with Lexical Features. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.
5. M. Joshi, S. Pakhomov, T. Pedersen, and C.G. Chute. A Comparative Study of Supervised Learning as Applied to Acronym Expansion in Clinical Reports. *AMIA Annual Symposium Proceedings*, 2006:399, 2006.
6. A Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the American Medical Informatics Association Symposium*, pages 17–21, 2001.
7. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.