Focus on i2b2 Obesity NLP Challenge JAMIA

Viewpoint Paper ■

Recognizing Obesity and Comorbidities in Sparse Data

Özlem Uzuner, PhD

Abstract In order to survey, facilitate, and evaluate studies of medical language processing on clinical narratives, i2b2 (Informatics for Integrating Biology to the Bedside) organized its second challenge and workshop. This challenge focused on automatically extracting information on obesity and fifteen of its most common comorbidities from patient discharge summaries. For each patient, obesity and any of the comorbidities could be Present, Absent, or Questionable (i.e., possible) in the patient, or Unmentioned in the discharge summary of the patient. i2b2 provided data for, and invited the development of, automated systems that can classify obesity and its comorbidities into these four classes based on individual discharge summaries. This article refers to obesity and comorbidities as *diseases*. It refers to the categories Present, Absent, Questionable, and Unmentioned as *classes*. The task of classifying obesity and its comorbidities is called the *Obesity Challenge*.

The data released by i2b2 was annotated for textual judgments reflecting the explicitly reported information on diseases, and intuitive judgments reflecting medical professionals' reading of the information presented in discharge summaries. There were very few examples of some disease classes in the data. The Obesity Challenge paid particular attention to the performance of systems on these less well-represented classes.

A total of 30 teams participated in the Obesity Challenge. Each team was allowed to submit two sets of up to three system runs for evaluation, resulting in a total of 136 submissions. The submissions represented a combination of rule-based and machine learning approaches.

Evaluation of system runs shows that the best predictions of textual judgments come from systems that filter the potentially noisy portions of the narratives, project dictionaries of disease names onto the remaining text, apply negation extraction, and process the text through rules. Information on disease-related concepts, such as symptoms and medications, and general medical knowledge help systems infer intuitive judgments on the diseases.

■ **J Am Med Inform Assoc.** 2009;16:561–570. DOI 10.1197/jamia.M3115.

Introduction

Narrative patient records allow doctors to write precise notes. The narratives do not contain controlled vocabularies, and thus allow doctors flexibility of expression.[1] However, the narratives also make information contained inaccessible to automated clinical systems. Natural language processing (NLP) and medical language processing (MLP) focus on technologies that can extract structured information from narratives.[2]

Affiliations of the author: University at Albany, SUNY, Albany, NY; Middle East Technical University, Northern Cyprus Campus, Kalkanli, Guzelyurt, Mersin 10, Turkey.

This work was supported in part by the NIH Road Map for Medical Research Grants U54LM008748. Institutional Review Board approval has been granted for the studies presented in this manuscript. The author thanks all participating teams for their contributions to the challenge, and AMIA for its support in organizing the workshop that accompanied the challenge.

Correspondence: Özlem Uzuner, College of Computing Information, University at Albany, SUNY, Draper 114A, 135 Western Ave, Albany, NY, 12222; e-mail: <ouzuner@albany.edu>.

Received for review: 12/22/08; Accepted for publication: 04/07/09.

The Obesity Challenge was motivated by the clinical need for technologies that can help counter the current obesity epidemic.[3] Its goal was to systematically evaluate NLP and MLP systems. Run as a shared task, the challenge was organized as a part of an i2b2 (Informatics for Integrating Biology to the Bedside) "Driving Biology Project." A total of 30 teams participated in the Obesity Challenge and met at a workshop cosponsored by the American Medical Informatics Association. This paper provides an overview of the challenge, describes the data and the evaluation metrics, reviews the best performing systems, and identifies directions for future MLP research.

Related Work

Systematic, head-to-head evaluations of technology can help advance state of the art and guide future research.[4] Shared tasks provide a way of conducting such evaluations. They provide the participants with a common set of training documents annotated with the ground truth for a particular task and evaluate all participants on the same held-out set.

Outside the medical domain, shared tasks have included the Message Understanding Conference[5] and the Text Retrieval Evaluation Conferences (TREC),[6] organized by the National

Institute of Standards and Technology.[7] Shared tasks for biomedicine have included BioCreAtIvE[8] and TREC Genomics.[9]

In 2006, we organized the first MLP shared task on clinical narratives.[10] This task focused on two challenges involving discharge summaries: automatic de-identification of personal health information (the De-identification Challenge)[11] and automatic evaluation of the smoking status of patients (the Smoking Challenge).[12] These shared tasks were followed by a similar effort of the University of Cincinnati Computational Medicine Center.[13] The Obesity Challenge continued i2b2's efforts to make existing clinical records available to the research community. Extracting information about obesity and comorbidities from narrative discharge summaries was the focus of this challenge.

Challenge Task: Recognition of Obesity and Comorbidities

To define the Obesity Challenge task, two experts from the Massachusetts General Hospital Weight Center studied 50 (25 each) random pilot discharge summaries from the Partners HealthCare Research Patient Data Repository. The experts identified fifteen frequently occurring obesity comorbidities: asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus (DM), gallstones/cholecystectomy, gastroesophageal reflux disease (GERD), gout, hypercholesterolemia, hypertension (HTN), hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency. They determined the Obesity Challenge task as automatic classification of obesity and the above comorbidities, referred to as *diseases*, as Present, Absent, or Questionable in a patient, or Unmentioned in the discharge summary of the patient. We define these *classes* as follows:

1. Present: the patient has/had the disease.
2. Absent: the patient does/did not have the disease.
3. Questionable: the patient may have the disease.
4. Unmentioned: the disease is not mentioned in the discharge summary.

We expect that the technologies developed in response to the challenge will be useful for indexing, classifying, and summarizing obesity-related facts found in discharge summaries. All relevant Institutional Review Boards approved the i2b2 Obesity Challenge.

Obesity Challenge Data

Data Draw and De-identification

Obesity Challenge data consisted of 1237 discharge summaries from the Partners HealthCare Research Patient Data Repository. These data were chosen from the discharge summaries of patients who were overweight or diabetic and had been hospitalized for obesity or diabetes sometime since 12/1/04. Some of the selected summaries included no mention of the stems "obes" and "diabet", others included at least one mention of these stems.

De-identification was performed semi-automatically. All private health information was replaced with synthetic identifiers.[11]

Annotation

Tasks

The data for the challenge were annotated by two obesity experts from the Massachusetts General Hospital Weight Center. The experts were given a *textual task*, which asked them to classify each disease (see list of diseases above) as Present, Absent, Questionable, or Unmentioned based on explicitly documented information in the discharge summaries, e.g., the statement "the patient is obese". The experts were also given an *intuitive task*, which asked them to classify each disease as Present, Absent, or Questionable by applying their intuition and judgment to information in the discharge summaries, e.g., the statement "the patient weighs 230 lbs and is 5 ft 2 inches". We refer to the textual task annotations as *textual judgments* and the intuitive task annotations as *intuitive judgments*.

Given the tasks, the experts agreed that:

- Textual judgments would require no reasoning.
- Intuitive judgments would generally agree with a textual Present, Absent, or Questionable judgment. The focus of the intuitive task would be on diseases marked Unmentioned.
- A textual judgment of Unmentioned, in the absence of information from the discharge summary supporting an inference about the disease, would translate to an intuitive judgment of Absent
- Information that would allow inference of diseases would include mentions of examination and test results, e.g., blood pressure or blood sugar measurements, physical characteristics, e.g., body mass index, and the medication and diseases discussed in the discharge summary.

Agreement and Tie-breaking

The two experts independently annotated our 1237 discharge summaries. The kappa (κ) agreement[14] between the two annotators on each disease is shown in Table 1. The lowest κ on textual judgments was 0.71. For 12 diseases, κ on textual judgments was above 0.8; for four diseases, κ on

Table 1 ■ Kappa Agreement on Textual and Intuitive Judgments

Comorbidity (Disease)	Textual Kappa	Intuitive Kappa
Asthma	0.90	0.76
Atherosclerotic CV disease (CAD)	0.78	0.81
Congestive heart failure (CHF)	0.81	0.74
Depression	0.92	0.86
Diabetes mellitus (DM)	0.91	0.87
Gallstones/cholecystectomy	0.92	0.90
GERD	0.89	0.59
Gout	0.93	0.92
Hypercholesterolemia	0.87	0.68
Hypertension (HTN)	0.82	0.67
Hypertriglyceridemia	0.71	0.72
Obesity	0.91	0.86
Obstructive sleep apnea (OSA)	0.92	0.92
Osteoarthritis (OA)	0.76	0.76
Peripheral vascular disease (PVD)	0.94	0.73
Venous insufficiency	0.79	0.44

CV = cardiovascular; GERD = gastroesophageal reflux disease.

Table 2 ■ Distribution of Classes between Textual and Intuitive Ground Truth

	Intuitive Present	Intuitive Absent	Intuitive Questionable	No Intuitive Class (No Agreement)
Textual Present	5021	2	0	377
Textual Absent	1	126	0	25
Textual Questionable	5	1	18	32
Textual Unmentioned	500	12327	20	1219
No Textual Class (No Agreement)	25	6	2	0

textual judgments was between 0.71 and 0.79. The lowest κ on intuitive judgments was 0.44. For seven diseases, κ on intuitive judgments was above 0.8; for six of the diseases, κ on intuitive judgments was between 0.6 and 0.79. Although the κ values are open to interpretation,[15] κ of 0.8 is widely used as the threshold for "almost perfect agreement", κ values of 0.6–0.79 indicate "substantial agreement"[14]. Please see the online supplement at http://jamia.org for a description of agreement calculation and extended analysis of agreement.

After annotation, a resident from the Massachusetts General Hospital resolved the disagreements in textual judgments. Majority vote among the three annotators determined the ground truth for the textual task. In the absence of a third obesity expert who could resolve the disagreements in intuitive judgments, only judgments agreed on by the two obesity experts were used as the ground truth for the intuitive task. Table 2 shows the correspondence between the ground truth textual and intuitive judgments. Most textual Present judgments map to intuitive Present judgments. Similar observations hold for the other classes.

Final Data

Table 3 and Table 4 show data distribution into training and test sets per disease. The distributions are non-uniform. In studying datasets with unbalanced class distributions, it is easier to focus on the better populated classes and ignore the less well-represented ones due to their limited contribution to overall performance. In our case, the less well-represented classes indicate the possibility or absence of a disease in a patient. Accurate recognition of these classes allows their inclusion in structured knowledge bases that can support future clinical decisions. Please refer to the online supplement at http://jamia.org for Table 5 and baseline results on these data.

Methods

We evaluated system performances using micro- and macro-averaged precision (P), recall (R), and F-measure (F_1). Given the emphasis of the Obesity Challenge on the less well-represented classes, we used macro-averaged F-measure as the primary metric for evaluation. Micro-averaged F-measure maintained a global perspective on the results.

Evaluation Metrics

For each disease, the macro-averaged metrics represent the arithmetic mean of the precision, recall, and F-measure on the Present, Absent, Questionable, and Unmentioned classes that are observed in the ground truth for that disease (see Eqs 1, 2 and 3). The macro-averaged precision, recall, and F-measure of the system are obtained from the precision, recall, and F-measure on the classes observed in the ground truth for all diseases. In these formulae, M is the number of classes.

Table 3 ■ Distribution of Textual Judgments into Training and Test Sets

	Present		Absent		Questionable		Unmentioned		Total	
Diseases	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Asthma	93	68	3	2	2	2	630	432	728	504
CAD	399	277	23	22	7	2	292	196	721	497
CHF	310	205	11	11	0	0	399	280	720	496
Depression	104	72	0	0	0	0	624	434	728	506
DM	485	338	15	12	7	3	219	150	726	503
Gallstones	109	87	4	2	1	0	615	418	729	507
GERD	118	69	1	1	5	1	599	433	723	504
Gout	90	52	0	0	4	0	634	453	728	505
Hypercholesterolemia	304	213	13	6	1	4	408	279	726	502
HTN	537	374	12	6	0	0	180	121	729	501
Hypertriglyceridemia	18	10	0	0	0	0	711	497	729	507
Obesity	298	198	4	3	4	3	424	289	730	493
OSA	105	69	1	0	8	2	614	432	728	503
OA	115	86	0	0	0	0	613	416	728	502
PVD	102	64	0	0	0	0	627	497	729	507
Venous insufficiency	21	10	0	0	0	0	707	497	728	507
Total	3208	2192	87	65	39	17	8296	5770	11630	8044

CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus; GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea; OA = osteo arthritis; PVD = peripheral vascular disease.

Table 4 ■ Distribution of Intuitive Judgments into Training and Test Sets

Diseases	Present		Absent		Questionable		Total	
	Training	Test	Training	Test	Training	Test	Training	Test
Asthma	86	68	596	403	0	0	682	471
CAD	391	272	265	185	5	1	661	458
CHF	308	205	318	229	1	4	627	438
Depression	142	105	555	372	0	0	697	477
DM	473	333	205	146	5	0	683	479
Gallstones	101	80	609	411	0	0	710	491
GERD	144	93	447	331	1	2	592	426
Gout	94	61	616	439	2	0	712	500
Hypercholesterolemia	315	242	287	189	1	0	603	431
HTN	511	358	127	88	0	0	638	446
Hypertriglyceridemia	37	25	665	461	0	0	702	486
Obesity	285	192	379	255	1	0	665	447
OSA	99	66	606	427	8	2	713	495
OA	117	91	554	367	1	4	672	462
PVD	110	65	556	399	1	1	667	465
Venous insufficiency	54	29	577	398	0	0	631	427
Total	3267	2285	7362	5100	26	14	10655	7399

CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus; GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea; OA = osteo arthritis; PVD = peripheral vascular disease.

Equation 1—Macro-averaged Precision (P_{macro})

$$P_{macro} = \frac{\sum_{i=1}^{M} P_i}{M}$$

Equation 2—Macro-averaged Recall (R_{macro})

$$R_{macro} = \frac{\sum_{i=1}^{M} R_i}{M}$$

Equation 3—Macro-averaged F-measure ($F_{1_{macro}}$)

$$F_{1_{macro}} = \frac{\sum_{i=1}^{M} F_{1_i}}{M}$$

Macro-averages give equal weight to each class, including rare ones.[16] As a result, two systems that make the same raw number of mistakes can end up with two different macro-averaged scores.

Equation 4 and Equation 5 show the formulae for computing micro-averaged precision and recall from true positives (TP), false positives (FP), and false negatives (FN) for each class.[16,17] In these formulae, M is the number of classes. Micro-averaged F-measure is the harmonic mean of micro-averaged precision and recall (Eq 6). Micro-averages give equal weight to each sample regardless of its class. They are dominated by those classes with the greatest number of samples.

Equation 4—Micro-averaged Precision (P_{micro})

$$P_{micro} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} TP_i + \sum_{i=1}^{M} FP_i}$$

Equation 5—Micro-averaged Recall (R_{micro})

$$R_{micro} = \frac{\sum_{i=1}^{M} TP_i}{\sum_{i=1}^{M} TP_i + \sum_{i=1}^{M} FN_i}$$

Equation 6—Micro-averaged F-measure ($F_{1_{micro}}$)

$$F_{1_{micro}} = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}}.$$

Significance Test

We determined the significance of the difference of the systems' performance using the Z test on two proportions.[18,19] We used a two-tailed test with a Z value of ± 1.645 and confidence level of 0.9.[20]

Obesity Challenge Submissions

A total of 30 teams participated in the Obesity Challenge (see Table 6). Training data were released in March 2008. Test data were released in June 2008. Each team submitted up to three system runs for predicting textual judgments and three for predicting intuitive judgments on test data.

We received a total of 68 textual and 68 intuitive system runs.[21–46] To obtain textual task results, we ranked each team on its best performing textual system run. To assess the intuitive task, we ranked each team on its best performing intuitive system run. We review the top ten textual and intuitive systems in ranked order below.

Top Ten Textual Systems

Of the top ten textual systems, Yang et al.,[22] Solt et al.,[42] Ware et al.,[28] Childs et al.,[24] Mishra et al.[43] Szarvas et al.,[21] and Deshazo et al.[26] filtered the narrative summaries from information indirectly related to the patient and marked negations and uncertainty through methods that resembled NegEx[47] or ConText.[48] In addition:

Yang et al. used a precompiled dictionary of disease, symptom, treatment, and medication terms. They looked for

Table 6 ■ Participating Teams

Team	Affiliations	Country
Ambert et al.	Oregon Health and Science University	United States
Barrett et al.	University of Victoria	Canada
Califf	Illinois State University	United States
Childs et al.	Lockheed Martin and SAGE Analytica	United States
DeShazo et al.	University of Washington	United States
Frunza et al.	University of Ottawa	Canada
Grabar et al.	LIPN–UMR 7030, Université Paris 13—CNRS Centre de Recherche des Cordeliers Université Paris Descartes HEGP AP-HP	France
Guillen	California State University, San Marcos	United States
Hara	Nara Institute of Science and Technology	Japan
Harkema et al.	University of Pittsburgh	United States
Ho et al.	IDI-NTNU	Norway
Jazayeri et al.	University of Alberta	Canada
Lan et al.	National University of Singapore Institute of Infocomm Research	Singapore
MacNamee et al.	Dublin Institute of Technology	Ireland
Mata et al.	Universidad de Huelva	Spain
Matthews	University of Edinburgh	Scotland
McInnes	University of Minnesota	United States
Medow	Boston University	United States
Meystre	University of Utah	United States
Mishra et al.	Centers for Disease Control and Prevention National Center for Public Health informatics	United States
Neves et al.	Centro Nacional de Biotecnología Universidad Complutense de Madrid	Spain
Patrick et al.	University of Sydney	Australia
Pedersen	University of Minnesota, Duluth	United States
Peshkin et al.	Harvard Alias-I, Inc.	United States
Savova et al.	Mayo Clinic MITRE	United States
Solt et al.	Budapest University of Technology and Economics TextMiner, Ltd, Budapest	Hungary
Szarvas et al.	University of Szeged	Hungary
Ware et al.	MedQuist West Virginia University	United States
Yang et al.	University of Manchester	UK

sentences with either exact or approximate matches. For documents that contained more than one sentence about a disease, they determined the class for that disease based on a weighted combination of the evidence in sentences.[22]

Solt et al. stripped the documents of personal identifiers, expanded abbreviations, and split discharge summaries into sections. To mark a disease as Present, they developed a rule-based classifier with disease names, synonyms, spelling variants, and semantically related terms. They partitioned text using contextual clues that indicate negative or uncertain statements and fed the partitions into a series of binary classifiers that determined whether each disease was Questionable, Absent, or Present, in that order. Diseases that failed to receive any of these three labels were labeled Unmentioned.[42]

Ware et al. used regular expressions with a set of disease-related keywords and their synonyms. They assumed that keywords not marked as negated, historical, or associated with a relative would indicate a disease is present.[28]

Childs et al. used the rule-based Rocket AeroText information extraction system[49] with keywords, their synonyms, and patterns generated by medical experts. They weighed and combined the evidence for each class of each disease.[24]

Mishra et al. marked the text with a set of disease-related keywords compiled by analyzing the training set. They determined the total number of positive, negative, and uncertain assertions for each disease in a discharge summary. The class with the highest number of assertions related to the disease labeled the disease. Ties were broken in favor of positive assertions.[43]

Szarvas et al. used term frequency and conditional probability in the Present class to preselect the most common terms that could aid classification. They supplemented this list with spelling variants and infrequent terms. The resulting dictionaries, along with disease contexts and document structure, formed the backbone of their rule-based system.[21]

Savova et al.[25] and Patrick et al.[44] deviated from the pattern of text filtering and negation extraction. Savova et al. com-

Table 7 ■ Micro- and Macro-averaged Results on Textual Judgments, Sorted by Macro-averaged F-Measure

	Macro-Averaged			Micro-Averaged		
Systems	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Yang et al.	0.8482	0.7737	**0.8052**	0.9723	0.9723	0.9723
Solt et al.	0.8318	0.7776	0.8000	0.9756	0.9756	0.9756
Ware et al.	0.8314	0.7542	0.7821	0.9718	0.9718	0.9718
Childs et al.	0.8169	0.7454	0.7762	0.9773	0.9773	**0.9773**
Mishra et al.	0.7485	0.8050	0.7718	0.9704	0.9704	0.9704
Szarvas et al.	0.7644	0.7600	0.7622	0.9729	0.9729	0.9729
Savova et al.	0.7701	0.7147	0.7377	0.9668	0.9668	0.9668
Patrick et al.	0.7971	0.6219	0.6737	0.9693	0.9693	0.9693
*Jazayeri et al.	0.7849	0.5779	0.6205	0.9514	0.9514	0.9514
†DeShazo et al.	0.8552	0.6240	0.6140	0.9639	0.9639	0.9639

Best F-measures are in bold. †System utilized external annotators.
*System description not available.

bined an information extraction system, a maximum entropy classifier, and an SVM. They evaluated these approaches, and determined the best one for each disease on each of the textual and intuitive tasks. They then allowed the identified best method to judge a disease for a task.[25]

Patrick et al. used a combination of rules and a decision-tree classifier with features that included signs, symptoms, and medication names related to each disease. They also leveraged the correlations between diseases.[44]

DeShazo et al. analyzed 300 of the discharge summaries, annotating them for information that supported ground truth textual judgments. They employed a rule base to propagate the information supporting ground truth judgments to the rest of the corpus.[26]

Top Ten Intuitive Systems

Most intuitive systems benefited from the output of the textual systems. Solt et al.[42] Szarvas et al.[21] and Childs et al.[24] determined a default mapping between textual and intuitive judgments and used it as the starting point. The top four intuitive systems employed rule-bases that incorporated "disease-specific, non-preventive medications and their brand names", disease-related procedures, and symptoms highly correlated with diseases,[42] "numeric expressions corresponding to measurements"[21], and medication names.[24,28]

Different from the top four, Ambert et al. took a machine learning approach to the intuitive task. They combined hot-spot filtering with error-correcting output codes. They identified words that demonstrated high information gain with respect to each disease, extracted the text within a 100-character window of these words, marked the negations, and vectorized the extracted text. Of the created vectors, "the ones that were absent any non-zero features" were automatically labeled Absent. The rest were labeled using error-correcting output codes that weighted each class inversely proportionally to its size.[45]

Meystre extracted sections and sentences from each discharge summary using regular expressions and rules. In these excerpts, he disambiguated acronyms and extracted concept identifiers from the Unified Medical Language System (UMLS).[50] He supplemented the identified concepts with medications and biomarker values that could indicate a disease. He determined intuitive labels using NegEx and ConText.[46]

Yang et al. based their intuitive predictions on evidence sentences containing information about symptoms, clinical measurements, and medications. They processed the sentences using clinical information, so the symptoms more directly related to a disease were more heavily weighted. The evidence sentences were considered to mark the presence of a disease unless a negation extractor marked them as negative or uncertain. In diseases with multiple evidence sentences, the information was combined.[22]

DeShazo et al. used SVMs for their intuitive system. This system used features derived from the text by the rule-based classifier they developed for the textual task.[26]

Table 8 ■ Significance Tests on the Top Ten Textual Systems

Systems	Solt et al.	Ware et al.	Childs et al.	Mishra et al.	Szarvas et al.	Savova et al.	Patrick et al.	Jazayeri et al.	†DeShazo et al.
Yang et al.	+*	*		*	*		*		
Solt et al.		*	*		*				
Ware et al.			+	+*	*	*	*		
Childs et al.				+	*				
Mishra et al.					+*	*	*		
Szarvas et al.							*		
Savova et al.							*		*
Patrick et al.									*
Jazayeri et al.									+

Sorted by macro-averaged F-measure. +marks pairs Not significantly different in macro-averaged F-measure. *marks pairs Not significantly different in micro-averaged F-measure. †System utilized external annotators. Only the upper diagonal is marked.

Table 9 ■ Micro- and Macro-averaged Results on Intuitive Judgments, Sorted by Macro-averaged F-Measure

Systems	Macro-Averaged			Micro-Averaged		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Solt et al.	0.7485	0.6571	**0.6745**	0.9590	0.9590	0.9590
Szarvas et al.	0.6999	0.6588	0.6727	0.9642	0.9642	0.9642
Childs et al.	0.7061	0.6540	0.6696	0.9582	0.9582	0.9582
Ware et al.	0.6410	0.6399	0.6404	0.9654	0.9654	**0.9654**
Ambert et al.	0.6383	0.6307	0.6344	0.9558	0.9558	0.9558
Meystre	0.6304	0.6387	0.6343	0.9566	0.9566	0.9566
Yang et al.	0.6383	0.6294	0.6336	0.9572	0.9572	0.9572
†DeShazo et al.	0.9722	0.6216	0.6292	0.9524	0.9523	0.9524
Matthews	0.6325	0.6256	0.6288	0.9509	0.9509	0.9509
Jazayeri et al.	0.6320	0.6257	0.6287	0.9508	0.9508	0.9508

Best F-measures are in bold. †System utilized external annotators.

Matthews evaluated as features stemmed word tokens, bigrams, trigrams, UMLS semantic types of concepts, and negation as extracted by NegEx. He identified the most useful features for each class and applied Bayesian networks to classify diseases.[33]

Obesity Challenge Results

The results for the textual task are shown in Table 7 and in Table 8. Table 7 shows that the best macro-averaged F-measure on the textual task was 0.8052; the best micro-averaged F-measure was 0.9773. Table 8 shows that the macro-averaged performance difference between the top two systems is not statistically significant. The top three systems are not significantly different in their micro-averaged F-measures. Table 9 and Table 10 show the top ten intuitive systems, as ranked by the macro-averaged F-measure. The best macro-averaged F-measure on the intuitive task is 0.6745; the best micro-averaged F-measure is 0.9654. Table 10 shows that the top three systems are not statistically different in either macro- or micro-averaged F-measures.

Table 11 shows that the top ten systems on the textual task had F-measures ranging from 0.92 to 0.97 on Present class. Their F-measures range from 0.97 to 0.99 on the Unmentioned class. On the Absent class, the F-measures range from 0.39 to 0.66; on the Questionable class, the F-measures range from 0 to 0.62. Table 12 shows that seven out of the top ten systems produced a zero F-measure on the Questionable class on the intuitive task. The best F-measure for this class is 0.12. The performance of the top ten systems on the Present class range from 0.92 to 0.95, while the top ten systems on the Absent class performed in a range from 0.97 to 0.98.

Discussion

Rule-based approaches played a significant role in the top ten systems in the textual task. Machine learning approaches contributed to the top ten systems in the intuitive task but were less dominant in the textual task.

Given the similar approaches taken by the top ten textual systems, we expect that their performance differences resulted from the accuracy of their negation extraction modules and the completeness of their dictionaries. The approaches taken by the intuitive systems were more varied. In general, clinical information, world knowledge, and information from the textual task benefited the top ten intuitive systems. A subset of the top ten textual and intuitive systems took advantage of medical experts, indicating the value of engaging medical professionals in system development.

A subset of the top ten textual and intuitive systems encodes expert knowledge in the form of hand-crafted rules and patterns, generated either through direct interactions with domain experts or through (laypersons') observations on the ground truth created by domain experts. "Expert knowledge is a combination of a theoretical understanding of the problem and a collection of heuristic problem-solving rules that experience has shown to be effective in the domain"[51].

Table 10 ■ Significance Tests on the Top Ten Intuitive Systems

	Szarvas et al.	Childs et al.	Ware et al.	Ambert et al.	Meystre	Yang et al.	†DeShazo et al.	Matthews	Jazayeri et al.
Solt et al.	+*	+*		*	*	*	*		
Szarvas et al.		+*	*						
Childs et al.				*	*	*	*		
Ware et al.				+	+	+	+	+	+
Ambert et al.					+*	+*	+*	+*	+*
Meystre						+*	+*	+*	+*
Yang et al.							+*	+*	+*
†DeShazo et al.								+*	+*
Matthews									+*

Sorted by macro-averaged F-measure. +marks pairs Not significantly different in macro-averaged F-measure. *marks pairs Not significantly different in micro-averaged F-measure. †System utilized external annotators. Only the upper diagonal is marked.

Table 11 ■ Top Ten Textual Systems on Individual Classes (Aggregate Over All Diseases)

	Present			Absent			Questionable			Unmentioned		
Systems	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Yang et al.	0.94	0.97	0.96	0.71	0.62	**0.66**	0.75	0.53	**0.62**	0.99	0.98	0.98
Solt et al.	0.96	0.97	0.96	0.63	0.63	0.63	0.75	0.53	**0.62**	0.99	0.98	**0.99**
Ware et al.	0.95	0.97	0.96	0.59	0.60	0.60	0.80	0.47	0.59	0.99	0.98	0.98
Childs et al.	0.96	0.97	**0.97**	0.75	0.55	0.64	0.57	0.47	0.52	0.99	0.99	**0.99**
Mishra et al.	0.95	0.96	0.96	0.62	0.63	0.63	0.44	0.65	0.52	0.99	0.98	0.98
Szarvas et al.	0.97	0.95	0.96	0.64	0.63	0.64	0.47	0.47	0.47	0.98	0.99	0.98
Savova et al.	0.95	0.94	0.95	0.74	0.52	0.61	0.41	0.41	0.41	0.97	0.98	0.98
Patrick et al.	0.95	0.96	0.96	0.69	0.31	0.43	0.57	0.24	0.33	0.98	0.98	0.98
Jazayeri et al.	0.91	0.93	0.92	0.59	0.29	0.39	0.67	0.12	0.20	0.97	0.97	0.97
†DeShazo et al.	0.94	0.95	0.95	0.50	0.57	0.53	1.00	0.00	0.00	0.98	0.98	0.98

Best F-measures per class are in bold. Sorted by macro-averaged F-measure. †System utilized external annotators.

However, such knowledge is limited to a closed-domain, narrowly defined task. Expert systems based on this knowledge, e.g., the hand-crafted systems developed for the Obesity Challenge, perform well when tested within the domain of their focus; however, they require some work to be adapted to new tasks and domains.

Despite the limitations on their generalizeability, MLP systems that can address the Obesity Challenge with near-human-level performance were developed within a three month period. Although starting from an existing system was preferred for the development of some systems, e.g.,[24,46] most, including two of the best systems[22,42] developed for the Obesity Challenge, were built from scratch.

The main complexity and difficulty of the Obesity Challenge, in contrast to past challenges[12,13] and most mainstream MLP work, came from the focus on less well-represented classes. The worst macro-averaged F-measures on the challenge were 0.2237 and 0.3358, in the textual and intuitive tasks respectively.

In particular, the textual Questionable class contained some discharge summaries that were incorrectly classified by all system runs. One such summary, marked Questionable for GERD, stated "The patient was continued on her PPI for GERD prophylaxis. ... required increasing her dosage of Nexium secondary to GERD-like symptoms."

Similarly, for the textual Absent class, no system runs could correctly predict the judgment for CAD in a discharge summary which stated, "no history of cancer or heart disease." In general, textual Absent judgment required careful study of the context where diseases are mentioned. For example, recognizing the absence of diabetes when a patient "had no further insulin requirement and was not a diabetic" requires correct interpretation of this text. Only a subset of the submitted system runs correctly classified this case.

The Present class was easier to predict. For example, all systems correctly labeled a discharge summary which stated "adult onset diabetes mellitus". However, even the Present class was not straightforward when the discharge summary failed to mention the disease by name. For example, a discharge summary about "ventral hernia" and "atrial fibrillation" that did not mention "coronary artery disease" or "cardiovascular disease" was judged Present for CAD. Only a subset of the submitted system runs predicted this textual judgment. Prediction of textual Present judgments was even more difficult in summaries using biomarkers or other related information to describe a disease. For example, none of the system runs submitted to the i2b2 challenge could correctly predict the ground truth judgment for obesity on the discharge summary that stated "The patient's admission weight was 106.2 kg. Her discharge weight was 100.7 kilograms", and "weight should be monitored daily."

The textual Unmentioned class was the easiest to predict. Most of these judgments were classified correctly by almost all the submitted system runs. Those textual Unmentioned judgments that could not be predicted correctly demonstrate peculiarities of data. For example, author's reading of the statement "The patient was an obese male" indicates a textual label of Present for obesity and disagrees with the ground truth label of Unmentioned.

Table 12 ■ Top Ten Intuitive Systems on Individual Classes (Aggregate over All Diseases)

	Present			Absent			Questionable		
Systems	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Solt et al.	0.95	0.92	0.93	0.96	0.98	0.97	0.33	0.07	**0.12**
Szarvas et al.	0.97	0.92	0.94	0.96	0.99	0.97	0.17	0.07	0.10
Childs et al.	0.96	0.91	0.93	0.96	0.98	0.97	0.20	0.07	0.11
Ware et al.	0.95	0.94	**0.95**	0.97	0.98	**0.98**	0.00	0.00	0.00
Ambert et al.	0.95	0.92	0.93	0.96	0.98	0.97	0.00	0.00	0.00
Meystre	0.91	0.96	0.93	0.98	0.96	0.97	0.00	0.00	0.00
Yang et al.	0.96	0.91	0.93	0.96	0.98	0.97	0.00	0.00	0.00
†DeShazo et al.	0.97	0.88	0.92	0.95	0.99	0.97	1.00	0.00	0.00
Matthews	0.94	0.90	0.92	0.96	0.98	0.97	0.00	0.00	0.00
Jazayeri et al.	0.94	0.90	0.92	0.96	0.98	0.97	0.00	0.00	0.00

Best F-measures per class are in bold. Sorted by macro-averaged F-measure. †System utilized external annotators.

Given the characteristics of the data and the observations on performance on the less well-represented classes, removing the emphasis from these classes would have made the Obesity Challenge much more mainstream and much more straightforward, but not trivial. Eighty-five percent of the systems in the intuitive task and 93% of the systems in the textual task achieved micro-averaged F-measures above 0.8. Two of the best performing systems from the Obesity Challenge are open source and can either be downloaded for local installations or utilized online.[52,53]

Conclusions and Implications for Future Research

The Obesity Challenge demonstrates the difficulty of differentiating textual judgments from intuitive ones. The overlap in information used by automated systems for identifying textual and intuitive judgments and the author's observations on the Obesity Challenge data indicate that textual judgments of domain experts may differ from textual judgments of lay persons. In other words, the annotators' domain knowledge may have led them to consider some inferred information as explicit. As a result, some judgments that could be considered intuitive by lay persons were found among the textual judgments.[54]

However, even with unclear boundaries between textual and intuitive judgments, the automated systems built by lay persons effectively extracted much useful information from discharge summaries. These systems performed best on the most factual and objective pieces of information. They experienced more difficulty arriving at conclusions only medical experts could infer. Most of the factual and objective pieces of information were identified by simple rule-based systems armed with dictionaries of terms and negation extraction modules. Machine learning approaches that studied the patterns in the textual judgments provided a beginning to correctly predicting intuitive judgments. We should emphasize that the relative performance of the systems is likely to change if we have much larger corpora for both training and testing. The unavailability of such corpora is likely to be the largest bottleneck for future progress in MLP.

References ■

1. Van Ginneken A, De Wilde M, Van Mulligen E, Stam H. Can data representation and interface demands be reconciled? Approach in orca. AMIA Annu Symp Proc; 1997:779–83.
2. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.
3. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med 2007 Jul 26;357(4):370–9.
4. Friedman C, Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. AMIA Annu Symp Proc; 1998:855–9.
5. Grishman R, Sundheim B. Message Understanding Conference-6: A brief history. 16th Conference on Computational Linguistics, COLING 1996, pp 466–71.
6. Sparck Jones K. Reflections on TREC. Inf Proc Manag 1995;31(3): 291–314.
7. NIST Available at: http://www.nist.gov/speech/tests/. Accessed: Aug 7, 2008.
8. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. BMC Bioinform 2005;6(S1).
9. Hersh W, Bhupatiraju RT, Corley S. Enhancing access to the Bibliome: The TREC genomics track. Medinfo 2004;11(2):773–7.
10. Uzuner Ö, Szolovits P, Kohane I. i2b2 workshop on natural language processing challenges for clinical discharge summaries. Available at: http://www.i2b2.org/NLP. Accessed: Oct 21, 2008.
11. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5): 550–63.
12. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge summaries. J Am Med Inform Assoc 2008;15(1):14–24.
13. Pestian JP, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. BioNLP, 2007, pp 97–104.
14. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
15. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002;35(2):99–110.
16. Yang Y, Liu X. A Re-examination of text categorization methods. Proceedings of the ACM SIGIR Conference on Research and Development in Inf Retrieval, pp 42–9, 1999.
17. Ozgur A, Ozgur L, Gungor T. Text categorization with class-based and corpus-based keyword selection. Lecture Notes in Computer Science. Springer-Verlag. 2005;(3733):607–616.
18. Osborn CE. Statistical Applications for Health Information Management, 2nd edn, Boston: Jones & Bartlett Publishers, 2005.
19. Z-test for Two Proportions. Available at: http://www.dimensionresearch.com/resources/calculators/ztest.html. Accessed: Jun 19, 2008.
20. Chinchor N. The Statistical Significance of the MUC-4 Results, McLean, VA: 4th Conference Mess Understand, 1992, pp 30–50.
21. Szarvas G, Farkas R, Almási A, et al. Semi-automated construction of decision rules to predict morbidities from clinical texts. J Am Med Inform Assoc 2009;16:601–5.
22. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of a disease status from clinical discharge summaries. J Am Med Inform Assoc 2009;16:596–600.
23. Guillen R. Identifying obesity and Co-morbidities from medical records. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
24. Childs LC, Taylor RJ, Simonsen L, et al. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. J Am Med Inform Assoc 2009:571–5.
25. Savova G, Clark C, Zheng J, et al. The Mayo/MITRE system for discovery of obesity and its comorbidities. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
26. DeShazo JP, Turner AM. Hands-on NLP: An interactive and user-centered system to classify discharge summaries for obesity and Related Co-morbidities. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
27. Califf ME. Combining rules and naïve bayes for disease classification. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
28. Ware H, Mullett CJ, Jagannathan V. Natural language processing framework to assess clinical conditions. J Am Med Inform Assoc 2009:585–9.
29. Hara K. Classifying narrative patient records without any external resources. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
30. Grabar N, Hamon T, Dart T. Term variation and semantics for document classification and detection of obesity and its Co-morbidities cases. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
31. Harkema H, Piwowar H, Amizadeh S, et al. A baseline system for the i2b2 obesity challenge. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.

32. MacNamee B, Kelleher JD, Delany SJ. Medical language processing for patient diagnosis using text classification and negation labelling. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
33. Matthews MP. Bayesian networks and the i2b2 obesity challenge. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
34. Frunza O, Inkpen D. Representation and classification techniques for clinical data focused on obesity and its co-morbidities. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
35. Peshkin L, Cano C, Carpenter B, Baldwin B. Regularized logistic regression for clinical record processing. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
36. Ho B, Nytrø Ø, Bassøe CF. NLP obesity challenge: Using clinical markers for EHR classification. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
37. Mata J, Maña MJ, Bermúdez JM, Cruz NP, Jiménez P. Handling negation in classification of clinical texts. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
38. McInnes BT. Using CuiTools to identify obesity and its Co-morbidities in discharge summaries. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
39. Neves M, Carazo JM, Pascual-Montano A. Botero: A SVM classifier for clinical text in the obesity domain. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
40. Pedersen T. Learning high precision rules to make predictions of morbidities in discharge summaries. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
41. Barrett N, Weber-Jahnke J. An introduction to MLP driven by the i2b2 challenge. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
42. Solt I, Tikk D, Gál V, Kardkovács ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. J Am Med Inform Assoc 2009;16:580–4.
43. Mishra N, Cummo D, Arnzen J, Bonander J. A rule-based approach for identifying obesity and its co-morbidities in medical discharge summaries. J Am Med Inform Assoc 2009:576–9.
44. Patrick J, Asgari P. A brief summary about the approach and explanation of the attributes of the developed system for i2b2 challenge. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
45. Ambert KH, Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. J Am Med Inform Assoc 2009;16:590–5.
46. Meystre SM. Detecting patients suffering from obesity and common comorbidities by analyzing narrative clinical text. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2008.
47. Chapman WW, Bridewell W, Hanbury P, et al. Algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:301–10.
48. Chapman W, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. BioNLP 2007, pp 81–8.
49. AeroText information extraction system. Available at: http://www.rocketsoftware.com/products/rocket-aerotext. Accessed: Apr 1, 2009.
50. Unified Medical Language System (UMLS). Available at: http://www.nlm.nih.gov/research/umls/. Accessed: Nov 21, 2008.
51. Luger GF. Artificial Intelligence: Structures and strategies for Complex Problem Solving, 6th edn, Boston: Addison-Wesley, Pearson Education, 2009.
52. Solt I. Automatic semantic annotation of medical discharge summaries. Available at: http://152.66.244.218/cgi-bin/demo.pl. Accessed: Apr 2, 2009.
53. Szarvas G. Automatic obesity-related morbidity identifier. Available at: http://www.inf.u-szeged.hu/rgai/?lang=en&page=obesity. Accessed: Apr 2, 2009.
54. Miller RA. Reference standards in evaluating system performance. J Am Med Inform Assoc 2002;9(1):87–8.