Looking for Subjectivity in Medical Discharge Summaries

The Obesity NLP i2b2 Challenge (2008)

Michael Roylance and Nicholas Waltner

Tuesday 3rd June, 2014

Paper

Journal of the American Medical Informatics Association Volume 16 Number 4 July / August 2009

Focus on i2b2 JAMIA Obesity NLP Challenge

Viewpoint Paper ■

Recognizing Obesity and Comorbidities in Sparse Data

ÖZLEM UZUNER, PHD

General Factoids

- The BioMedical field is awash in data.
- It is argued that up to 70% of important data about a patient is stored in *largely unstructured free text fields*[1]
- Although local hospitals like Swedish have heads of Informatics, there is still an active debate over how much machine learning can do to accurately diagnose patient using textual approaches.
- In spite of its enormous success in *Jeopardy!*, IBM's Watson has yet to make expected inroads in field medicine, although may well as Watson is distributed to mobile devices.
- Maybe the human doctors are the obstacle or maybe not?

[1] Please see: Shah, Stanford University. http://med.stanford.edu/ism/2013/april/clinical-notes.html#sthash.Gb42nykc.dpuf.

Task

- We worked on a medical dataset consisting of 1,237 patient discharge summaries used in the Obesity Challenge.
- Along with Obesity each patient was evaluated for an additional 15 co-morbidities such as Hypertension, Diabetes, Heart Disease, etc.
- Each patient's record was annotated using *textual* and *intuitive* classifications.
- The diseases were judged to be either Present, Absent, Questionable or Unmentioned for each patient.
- This led to a training corpus with 22,285 cases and a test one with 15,443.

Data Set - Textual Judgements

Table : Distribution of Textual Judgements 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
Diabetes	485	338	15	12	7	3	219	150	726	503
GERD	118	69	1	1	5	1	599	433	723	504
Gallstones	109	87	4	2	1	0	615	418	729	507
Gout	90	52	0	0	4	0	634	453	728	505
Hypercholesterolemia	304	213	13	6	1	4	408	279	726	502
Hypertension	537	374	12	6	0	0	180	121	729	501
Hypertriglyceridemia	18	10	0	0	0	0	711	497	729	507
OA	115	86	0	0	0	0	613	416	728	502
OSA	105	69	1	0	8	2	614	432	728	503
Obesity	298	198	4	3	4	3	424	289	730	493
PVD	102	64	0	0	0	0	627	443	729	507
Venous.Insufficiency	21	10	0	0	0	0	707	497	728	507
Total	3,208	2,192	87	65	39	17	8,296	5,770	11,630	8,044

Notes: 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.

Data Set - Intuitive Judgements

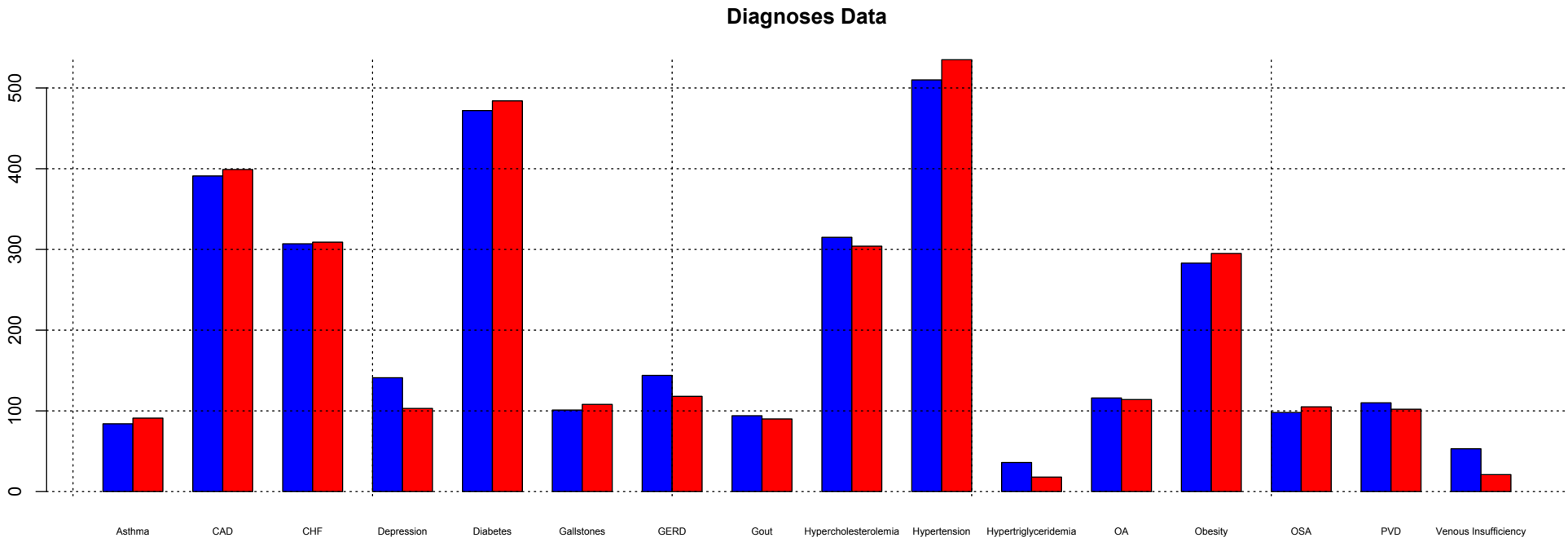
Table : Distribution of Intuitive Judgements into Training and Test Sets

	Present		Absent		Questionable		Unmentioned		Total	
Diseases	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Asthma	86	68	596	403	0	0	0	0	682	471
CAD	391	272	265	185	5	1	0	0	661	458
CHF	308	205	318	229	1	4	0	0	627	438
Depression	142	105	555	372	0	0	0	0	697	477
Diabetes	473	333	205	146	5	0	0	0	683	479
GERD	144	93	447	331	1	2	0	0	592	426
Gallstones	101	80	609	411	0	0	0	0	710	491
Gout	94	61	616	439	2	0	0	0	712	500
Hypercholesterolemia	315	242	287	189	1	0	0	0	603	431
Hypertension	511	358	127	88	0	0	0	0	638	446
Hypertriglyceridemia	37	25	665	461	0	0	0	0	702	486
OA	117	91	554	367	1	4	0	0	672	462
OSA	99	66	606	427	8	2	0	0	713	495
Obesity	285	192	379	255	1	0	0	0	665	447
PVD	110	65	556	399	1	1	0	0	667	465
Venous.Insufficiency	54	29	577	398	0	0	0	0	631	427
Total	3,267	2,285	7,362	5,100	26	14	0	0	10,655	7,399

Notes: 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.

Textual and Intuitive Counts

- The textual data is lumpy with the top four diseases (Hypertension, Diabetes,CAD (Coronary-Arterial) and Hypercholesterolemia) account for more than 50% of the data.
- Low frequency cases could cause classification confusion.



Data Set - A Quick Look

- Uzner reports high agreement kappa (κ) levels between annotators.
- The textual and intuitive diagnoses generally agreed quite well except for Depression, GERD, Hypertriglyceridemia and Venous Insufficiency.

Table : Agreement and Correlation between Textual and Intuitive Datasets

Diseases	**Textual κ**	**Intuitive κ**	**Correlation**
Asthma	0.90	0.76	0.919
CAD	0.78	0.81	0.928
CHF	0.91	0.74	0.858
Depression	0.92	0.86	**0.748**
Diabetes	0.91	0.87	0.926
GERD	0.92	0.90	**0.763**
Gallstones	0.89	0.59	0.956
Gout	0.93	0.92	0.885
Hypercholesterolemia	0.87	0.68	0.851
Hypertension	0.82	0.67	0.808
Hypertriglyceridemia	0.71	0.72	**0.523**
OA	0.91	0.86	0.815
OSA	0.92	0.92	0.933
Obesity	0.76	0.76	0.872
PVD	0.94	0.73	0.907
VenousInsufficiency	0.79	0.44	**0.473**
Averages	0.87	0.76	0.820

Competition Results

30 teams submitted results...textual macro-average F-scores were between 0.61 and 0.80 for the top ten teams.

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

Competition Results

30 teams submitted results...intuitive results were lower at 0.63 to 0.67, as one might expect.

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

	Macro-Averaged			Micro-Averaged		
Systems	**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

Take Aways

What did we learn from the paper:

- Most of the team did not rely super-heavily on pure ML, rather rule building on "standard language" seem to dominate the systems along with a lot of work on the naming of various diseases, etc.
- Intuitive judgements seem to be harder to machine learning (not so surprising).
- Each patient was diagnosed with 4.36 diseases - are the diseases similar or is there confusion?
- Possibly, sentiment measures could improve over a baseline, especially in areas where there was not strong agreement between textual and intuitive annotation, i.e. the human knew something that was not obvious in the text or vice versa.

Methodology

We obtained the dataset from i2b2 organization in XML format.

- Built a MySql database to house the data and build various tables around the data.
- Basic scrubbing and ETL (Extract, Transform and Load) was performed in Python and Perl.
- Used the Stanford Parser for POS tagging.
- Classification was done using Mallet andSKLearn (very handy especially with micro- and macro-averaging).
- Established a two class baseline (Present and Absent) and then added sentiment/subjectivity features.

Comp Ling Issues

As Gina pointed out in Week 6, "biomedical texts are not really English" !!!!

- POS X comes up nearly 30% of the time.
- Punctuation is very heavy owing to abbreviations.

Table : Part of Speech Counts

POS	Count	Percentage	POS	Count	Percentage
X	354,165	28.4	CC	28,902	2.3
NN	198,815	15.9	VBN	28,441	2.3
PUNC	147,095	11.8	RB	28,031	2.2
NNP	124,185	9.9	VB	20,515	1.6
JJ	93,352	7.5	PRP	18,060	1.4
IN	91,270	7.3	TO	17,915	1.4
CD	66,893	5.4	VBZ	16,474	1.3
DT	54,860	4.4	PRP$	12,653	1.0
VBD	46,635	3.7	VBP	10,895	0.9
NNS	46,234	3.7	VBG	9,972	0.8

Results

Sentiment and subjectivity features in many cases lowered classification accuracy. However, notable gains were found in the intuitive categories.

Table : Classification Results

Category	Sub-Task	Micro/Macro Intuitive	Micro/Macro Textual	Comment
Base Line	Uni-gram without StopWords	47.6 / 83.1	51.6 / 87.1	
	without X POS	39.1 / 72.5	40.9 / 73.1	
	without X -LBR- -RRB . , etc	39.1 / 72.5	40.9 / 73.1	
POS Tags	Pronouns-only	47.4 / 82.4	51.3 / 87.0	
	Nouns-only	47.4 / 82.1	50.2 / 84.7	
	Verbs-only	45.0 / 76.6	48.5 / 84.4	
	Adjectives-only	46.6 / 80.5	49.6 / 85.0	
	Adverbs-only	47.2 / 78.9	50.5 / 85.7	
	Adjectives and Adverbs-only	45.6 / 75.9	49.3 / 83.0	
	All Tags	47.9 / 80.2	51.0 / 86.0	
Polarity	Simple (positive/negative counts)	48.0 / 80.2	51.0 / 56.0	
	Complex (positive weak, positive strong)	47.2 / 82.6	51.3 / 86.8	
Combinations	Simple Polarity without X	39.2 / 73.2	40.8 / 72.3	
	Complex Polarity without X	39.5 / 71.8	40.8 / 71.6	
Other	Unique Words per Diagnosis	46.4 / 65.1	46.6 / 69.9	
	Highest Probability Words per Diagnosis	46.1 / 76.5	48.2 / 74.7	

Initial Conclusions

Did we fail or is something else going on?

- It may simply be the case that medical literature is largely absent emotive descriptions of patient discharge summaries.
- Alternatively, it may simply be the case that standard lexicons of subjectivity are insufficient for the medical domain.
- However, it is clear that there is a high degree of correlation between the various diseass.
- Hence, a more interesting question might be too ask whether are there fundamental drivers underneath these 16 diseases?
- Perhaps, unsupervised machine learning techniques can shed further light on what we already know?

An Unsupervised Approach

Both cluster and principal component analysis indicate that there is a higher structure to the co-morbidity data. PCA indicates that five-factors explain 50% of the variance in patient diagnoses...

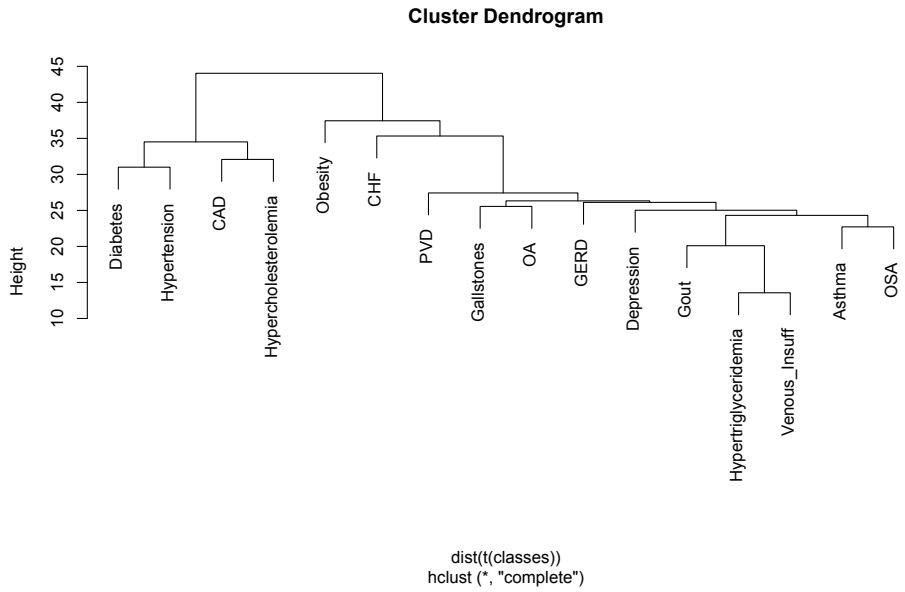


Figure : Simple Clustering

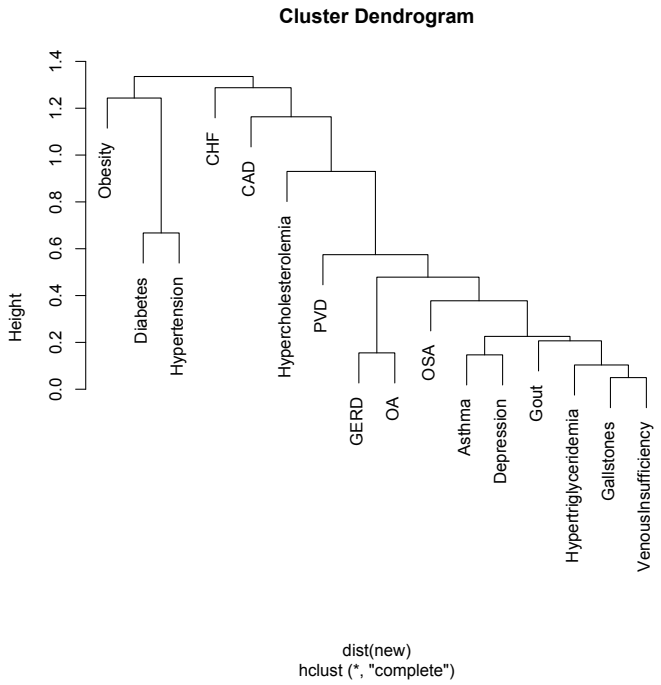


Figure : Five Factor PCA Model

Final Write-Up

Further items to research:

- Can combining both textual and intuitive features provide a better basis for diagnosis?
- Can other features be added to improve subjectivity accuracy?
- Can a decision tree be developed to arrive in the most likely disease cluster versus ending up with multiple diagnoses?