

Simple approaches to disease classification based on clinical patient records

György Szarvas, Richárd Farkas, Attila Almási, Veronika Vincze,

Research Group on Artificial Intelligence of the Hungarian

Academy of Sciences and University of Szeged

H-6720 Szeged, Aradi vértanúk tere 1., Hungary

István Hegedûs, Róbert Busa-Fekete, Róbert Ormándi

University of Szeged, Department of Informatics

H-6720 Szeged, Árpád tér 2., Hungary

Abstract

*In this study we describe the system submitted by the team of University of Szeged to the Second I2B2 Challenge in Natural Language Processing for Clinical Data. The challenge focused on the development of automatic systems that addressed the following question: **Who's obese and what co-morbidities do they (definitely / most likely) have?** using clinical documents. Target diseases included **obesity** and its 15 most frequent co-morbidities exhibited by patients, while the target labels corresponded to expert judgements based on **textual evidence** and **intuition** (separately). Our approach exploited statistical methods to pre-select the most common (and most confident) terms and evaluated outlier documents by hand to discover infrequent terms and spelling variants. We expected a system with dictionaries gathered semi-automatically to show a good performance with moderate development costs (we examined just a small proportion of the patient records manually). Our best submission achieved an accuracy score of 97.29% for classification based on textual evidence (macro-average $F_{\beta=1} = 76.22\%$) and an accuracy score of 96.42% for intuitive judgements (macro-average $F_{\beta=1} = 67.27\%$).*

INTRODUCTION

The *Obesity Challenge* organised by the Informatics for Integrating Biology and the Bedside (I2B2), a National Center for Biomedical Computing, asked participants to construct systems that could correctly replicate the textual and intuitive judgments of the medical experts on obesity and co-morbidities based on narrative patient records. Since hospitals usually store a considerable amount of information (patient data) as free text, such systems have a great potential in aiding research on obesity due to their capability to

process large document repositories both cost and time efficiently. Previous evaluation campaigns for clinical text classification include [1] and [2].

Task description

The target diseases included *obesity* and its 15 most frequent co-morbidities exhibited by patients, while the target labels corresponded to expert judgements based on *textual evidence* and *intuition*. That is, for each patient, both what the text explicitly said about obesity and co-morbidities, and what the text implied about obesity and co-morbidities, were provided as gold standard labels by obesity experts.

The dataset consisted of 1237 discharge summaries. Each document had been annotated for obesity and the other 15 diseases. Out of these documents, 730 were made available to the challenge participants for development and the remaining 507 documents constituted the evaluation set. In the textual annotation part, cases of disagreement in labeling were resolved by a resident doctor. In the intuitive annotation part there was no tie-breaking so documents with inconsistent labeling were simply excluded from the training and evaluation data for that particular disease. Those few documents where the third annotator could not decide on a final label for textual annotation were also discarded (for that disease). This meant that the number of training and test examples varied from disease to disease, especially for the intuitive task – and the third annotator could not decide on a final textual label for about 1% of the documents.

For a more comprehensive description of the task and the data itself, see www.i2b2.org/NLP/.

Our approach

Our approach focused on the rapid development of dictionary-lookup-based systems, which also took

into account the document structure and the context of disease terms for classification. To achieve this, we used statistical methods to pre-select the most common (and most confident) terms and evaluated outlier documents by hand to discover infrequent terms and spelling variants. Uncertainty and negation detection exploited keyword lists to identify negations/hedges and delimiter lists to determine their scope. Terms within the scope of a negation or uncertain cue were handled with respect to this information. We expected a system with dictionaries gathered semi-automatically to show a good performance with moderate development costs (we examined just a small proportion of the patient records manually).

METHODS

Textual model

For the challenge we applied a dictionary lookup-based system. That is, we collected a dictionary of terms and abbreviations for each disease separately, processed each document and collected occurrences of dictionary terms from the text. Sentences containing disease terms were then further evaluated to decide the appropriate class label for the corresponding disease. Further evaluations included a judgement of relevance (information on the patient and not on family members, etc.) and an analysis of context to detect negation/uncertainty.

After locating and evaluating all the relevant pieces of information in the document, the main decision function of our system was based on the following rules:

1. classify a document as **yes** if any terms were matched in an assertive context
2. classify a document as **no** if any terms were matched in a negative context /and was not classified in previous step/
3. classify a document as **questionable** if any terms were matched in an uncertain context /and were not classified in a previous step/
4. classify as **unmentioned** if none of the previous steps triggered a different labeling

Intuitive model

Our intuitive model was based on the textual model. That is, we attempted to discriminate the documents classified as **unknown** by our textual classifier to **yes** or **no** classes. When the textual system assigned a label that was different from

unknown, we accepted that decision as an intuitive judgement as well. Obviously this assumption is somewhat simplistic, but based on our observations on the training dataset, this assumption turned out to be quite reasonable.

In order to classify textual **unmentioned** documents, we collected phrases and numeric expressions which indicated an intuitive **yes** label. Such phrases were typically names of associated drugs and medication, or phrases related to certain social habits of the patients (e.g. cigarette for hypertension), while numeric expressions were tension values, weight, etc. Since these terms usually contained implicit information on the corresponding disease, it made no sense to evaluate their context for hedge cues. That is, these lists were not used to predict intuitive **questionable** labels.

After locating and evaluating all relevant pieces of information in the document, the main decision function of our system was based on the following rules:

1. Classify textual **yes/no/questionable** accordingly
2. For textual **unmentioned** docs:
 - (a) classify a document as an intuitive **yes** if any intuitive-terms were matched in an assertive context
 - (b) classify a document as an intuitive **yes** if a numeric expression was below/above the pre-defined threshold
 - (c) classify a document as an intuitive **no** if none of the previous steps triggered a different labeling

System components

Keyword / Excluding term selection The terms included in the dictionary were gathered semi-automatically: we ranked each term according to their frequency + positive class (**yes**) conditional probability scores. The top ranked terms were considered for addition to a disease-name dictionary manually. This way a 95% complete dictionary could be gathered quite rapidly. For each of the 16 diseases there were some **yes** documents that were not captured this way, these were examined manually for potential disease terms that were too infrequent to capture by statistical methods (e.g. $freq(hyper\ tg)=2$; $freq(gallbladder\ stone)=1$...

Pseudo terms (longer phrases containing a previously added disease term) were then collected by

a similar method to avoid the overfitting of dictionary lookup (e.g. *depression*, but not *st depression* or *hypertension* but not *pulmonary hypertension*). Collected disease name dictionaries were then extended with a few spelling variants manually, to handle different spellings of the same term.

Irrelevant contexts (U-dictionary) We also made use of an *unknown* dictionary that triggered the exclusion of the text from further processing. This way we neglected sections under headings like *FAMILY HISTORY*: and also phrases like *son with...*, *family history of...*

Negation / Uncertainty detection The system with the above-described components was able to tag docs with **yes** labels or left them as **unmentioned**. Doing this, we also extracted sentences with disease names from Y-tagged **questionable** & **no** docs and these sentences served as the basis for implementing a simple negation/uncertainty detection module. This exploited a list of negation / uncertainty cues and a list of delimiters (that triggered the end of scope). Hedge, negation cues and delimiter words/punctuation were chosen so as to provide an optimal performance on the training dataset in terms of a macro-averaged F measure. That is, we selected each word from the extracted sentences that seemed to be meaningful delimiters or keywords, and discarded those that lowered performance automatically. This approach is similar to NegEx [3] and our biomedical text corpus annotated for negation and uncertainty [4] also demonstrates that this simple scope resolution approach works well for clinical texts.

The few **questionable** and **no** cases that were not covered this way were examined manually to extract such terms as *normotensive* for N-hypertension for example (or were neglected if we found no clear evidence for the Q/N label, and if extending dictionaries to capture the particular instance actually caused overfitting and errors on other examples).

Intuitive terms We extended the system with *intuitive dictionaries* that triggered intuitive **yes** labels. These *intuitive dictionaries* were used to classify a documents as an intuitive **yes** when it was judged to be **unmentioned** by the textual classifier system.

- *MedLine Plus*: These terms (typically names of associated drugs and medication, etc) were collected from the MedlinePlus encyclopedia and then filtered for intuitive positive class-

conditional probability.

- *C4.5*: We also extracted terms like these by training decision trees to discriminate intuitive **yes** and **no** documents using a vector space model representation of the documents. We made the assumption here that complex rules represented by decision trees of depth greater than 1 were unlikely to provide a meaningful classification result so we extracted the words from the nodes of the learnt decision trees and automatically evaluated them as single terms in our dictionary lookup system. Those terms were kept in the submitted system which improved the performance as single terms on the training dataset.

Numeric expressions We also added a model that looked for numeric expressions preceding or following certain keywords to classify intuitive **yes** documents. These were typically for obesity /weight/, hypertension /high tension values/, etc...). Thresholds for the numeric expressions were set to provide the optimal performance on the training dataset.

Example: *if the phrase 'ejection fraction' is found and the associated value is below 50, predict intuitive yes label for congestive heart failure.*

Unsuccessful approaches We also experimented with various other approaches to achieve comparable results or to fine-tune the above-described systems which seemed to provide no benefit even on the training dataset. These were:

- *Train classifiers to discriminate textual unmentioned and yes documents.* We used a simple vector space model representation of the training dataset but got worse results compared to a simple dictionary lookup system (with the dictionaries collected by using the above discussed procedure). Our previous work on the similar ICD-9 coding problem showed that it is possible to automatically replicate these dictionary lookup-based systems by training classifiers to extract the terms for the dictionaries [5]. We had no time to evaluate this precisely here.
- *Derive disease-disease relationships based on the assigned textual labels.* Since the diseases addressed in the challenge were the most frequent co-morbidities of obesity, we expected to find strong relationships between certain diseases. That is, we used the other 15 textual labels as

features and trained classifiers to predict the label of the 16th disease. We experimented with several classifiers here, but without success.

- *Delete sentences with a disease term and add the textual **yes** documents for subsequent intuitive **yes/no** training.* We had the idea that, besides containing explicit information related to a certain disease, documents should also contain implicit information related to it as well. Hence we discarded the sentences containing a disease term from textual **yes** documents and added these records as **yes** examples for subsequent intuitive classification. Alas, this approach did not provide any particular benefit compared to the straightforward method we applied (discriminate textual **unmentioned** documents to intuitive **yes** and **no** classes).

Unimplemented methods We had some ideas that were not implemented due to the challenge deadlines. These included the following:

- *Edit distance-based fuzzy-matching of disease terms.* Since medical narrative texts frequently contain misspellings, we considered experimenting with matching spelling variants with 1 or 2 edit distances from a given disease term.
- *Incorporate discarded intuitive documents to train intuitive **yes/no** classifiers.* We thought about adding those train documents that had no label provided as intuitive **yes** documents. As the most obvious cause of discarding a document from intuitive annotation was that one expert judged it a **yes** and the other a **no** document, it was likely that some (or many) of these still held some information that was relevant to that particular disease. However these additional documents would have provided quite noisy data so in the end we decided not to pursue this idea further.

RESULTS

Few words about the 3 uploads:

- *upload1:* the system developed and found best on train set (described above)
- *upload2:* uncertainty/negation dictionaries enriched by a linguist as she found appropriate (even though this lowered the F-macro on the train set)
- *upload3:* those few cues that caused the major part of the decrease in F-macro performance were deleted from the upload2 system (sg. in between upload1 and upload2).

According to the official evaluation, *upload1* gave the best performance on the test set, as expected from the stats on the train set. This was good for an F-macro of 84% on the train for our best model (degraded to 76% on test), and an intuitive F-macro of 82% on train (degraded to 67% on test). This system came 6th in the textual F-macro ranking and 2nd in the intuitive F-macro ranking. The micro-averaged results were in the high 90s as the system was especially accurate on the **yes** and **unmentioned** classes (**yes** and **no** for intuitive judgement), and these classes had much more examples than the **questionable** and textual **no** classes.

Detailed results

Details on the performance gain by each component, per class results on the test dataset and confusion matrices of our best submission can be found in page 5 (Tables 1-5).

DISCUSSION

Our intuitive model was based on the textual model. This is why we encountered worse performance in intuitive **questionable** tagging on the test data; we neglected textual **unmentioned** documents that got an intuitive **questionable** label because there were too few of them in the training data to model this phenomenon, especially without background medical knowledge). The figures show that we suffered greatly from the train/test distribution of **questionable** examples (9 **tU-iQ** docs and 17 **tQ-iQ** docs in train, while 11 and only 1 in the test set). There were 2 test documents that had no textual gold standard label (the third annotator was unable to resolve disagreement on the textual label) but got an intuitive **questionable** label, which we found odd.

The model suffered from a lack of coverage for the **no** and **questionable** classes in textual annotation as well (performance dropped from 84% to 76% in the textual task, mainly due to more **no** & **questionable** docs left as **unmentioned** than in the training set).

CONCLUSIONS

As regards the classification accuracy scores, the method proposed here looks quite promising for automated processing of large datasets to gather information on obesity and related diseases. Those classes that had a few hundred training examples (**yes** & **unmentioned** for textual and **yes** & **no** for intuitive annotation) generally achieved an F measure of about 97%. This suggests that

our approach is indeed capable of locating the most relevant pieces of information for each of the 16 diseases addressed in almost all of the documents. Lower scores were observed for infrequent classes (with only 1-10 examples on average per class/disease pair) and we think that having more examples for **questionable** cases and negative examples (textual **no** label) would result in a substantial improvement in performance on these particular classes as well. Overall, we believe that our results prove the feasibility of our approach and that even very simple systems with a shallow linguistic analysis can achieve remarkable accuracy scores for classifying clinical records.

In a wider context, the outstanding accuracy scores (micro-averaged results) here demonstrate the potential in Text Mining in the clinical domain.

Acknowledgements

The authors would like to thank the organizers of the I2B2 Obesity challenge and the annotators of the challenge dataset for their efforts that enabled us to work on this interesting and challenging problem. This work was supported in part by the NKTH grant of the Jedlik Ányos R&D Programme 2007 of the Hungarian government (codename TUDORKA7).

Address for Correspondence

György Szarvas, University of Szeged, Department of Informatics, Árpád tér 2., H-6720 Szeged, Hungary
szarvas@inf.u-szeged.hu

References

- [1] Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24, January 2008.
- [2] John P. Pestian, Chris Brew, Pawel Matykievicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of BioNLP Workshop of ACL*, pages 97–104, Prague, Czech Republic, 2007.
- [3] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 5:301–310, 2001.
- [4] György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of BioNLP Workshop of ACL*, pages 38–45, Columbus, Ohio, 2008.
- [5] Farkas Richárd and Szarvas György. Automatic construction of rule-based icd-9-cm coding systems. In *BMC Bioinformatics, Volume 9 Suppl 3*, 2008.

System	Train		Test	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Upload1	97.91	83.94	97.29	76.22
Upload1 w/o U-dict	97.57	82.07	96.88	73.10
Upload1 w/o neg/unc	97.26	51.23	96.81	51.03
Upload1 w/o both	96.93	51.03	96.47	50.82

Table 1: Textual results.

System	Train		Test	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Upload1	97.11	82.32	96.42	67.27
Upload1 w/o I-terms	96.21	81.57	95.42	66.42
Upload1 w/o numexp	96.90	82.15	96.26	67.13
Upload1 w/o both	96.00	81.39	95.26	66.28

Table 2: Intuitive results.

Disease	Textual		Intuitive	
	F_{micro}	F_{macro}	F_{micro}	F_{macro}
Asthma	98.81	82.47	98.73	97.42
CAD	91.15	85.13	93.89	62.57
CHF	93.15	77.81	93.84	62.83
Depression	98.42	97.16	97.48	96.61
Diabetes	95.43	81.60	96.66	96.03
Gallstones	98.82	79.06	99.19	98.48
GERD	98.41	73.34	91.78	57.80
Gout	99.21	97.93	99.20	98.18
Hypercholesterolemia	96.61	84.95	91.42	91.47
Hypertension	97.21	80.06	96.19	94.61
Hypertriglyceridemia	98.82	78.27	98.97	94.41
OA	96.81	94.23	93.51	59.93
Obesity	96.96	48.83	97.54	97.49
OSA	99.20	65.87	99.60	88.34
PVD	98.42	96.81	96.99	62.81
Venous Insufficiency	99.01	89.75	96.02	78.22

Table 3: Per disease results on test set.

	Train				Test			
	Y	N	Q	U	Y	N	Q	U
YES	3072	6	1	129	2089	7	2	94
NO	11	66	0	10	9	41	1	14
QUESTIONABLE	8	0	22	9	6	0	8	3
UNMENTIONED	46	18	5	8227	60	16	6	5688

Table 4: Textual confusion matrix.

	Train			Test		
	Y	N	Q	Y	N	Q
YES	3020	243	4	2096	186	3
NO	47	7315	0	61	5037	2
QUESTIONABLE	5	10	11	3	10	1

Table 5: Intuitive confusion matrix.