Semi-automated construction of decision rules to predict morbidities from clinical texts

Address for Correspondence

Richárd Farkas, rfarkas@inf.u-szeged.hu

Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and

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

AUTHORS:

Richárd Farkas

Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and

University of Szeged,

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

Tel.: +3662546714

e-mail: rfarkas@inf.u-szeged.hu

György Szarvas, PhD

Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and

University of Szeged, AND

Technische Universität Darmstadt, Department of Computer Science, Ubiquitous Knowledge

Processing Lab

Hochschulstr. 10

64289 Darmstadt, Germany

Tel.: +49615116 5430

e-mail: szarvas@tk.informatik.tu-darmstadt.de

István Hegedűs

University of Szeged, Department of Informatics,

H-6720, Szeged, Árpád tér 2., Hungary

e-mail: ist.hegedus@gmail.com

Attila Almási

University of Szeged, Department of Informatics,

H-6720, Szeged, Árpád tér 2., Hungary

Tel.: +3662544669

e-mail: vizipal@gmail.com

Veronika Vincze

University of Szeged, Department of Informatics,

H-6720, Szeged, Árpád tér 2., Hungary

Tel.: +3662546720

e-mail: vinczev@inf.u-szeged.hu

Róbert Ormándi

University of Szeged, Department of Informatics,

H-6720, Szeged, Árpád tér 2., Hungary

Tel.: +3662546714

e-mail: ormandi@inf.u-szeged.hu

Róbert Busa-Fekete, PhD

Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and

University of Szeged, AND

LAL, University of Paris-Sud, CNRS,

91898 Orsay, France

Tel.: +33164468361

e-mail: busarobi@lal.in2p3.fr

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 analyzed clinical discharge summary texts and addressed the following question: *'Who's obese and what co-morbidities do they (definitely / most likely) have?'*. Target diseases included ***obesity*** and its 15 most frequent co-morbidities exhibited by patients, while the target labels corresponded to expert judgments based on ***textual evidence*** and ***intuition*** (separately).

We applied statistical methods to pre-select the most common and confident terms and evaluated outlier documents by hand to discover infrequent spelling variants.

We expected a system with dictionaries gathered semi-automatically to have a good performance with moderate development costs (we examined just a small proportion of the records manually). Our submission achieved a micro-average $F_{\beta=1}$ score of 97.29% for classification based on textual evidence (macro-average $F_{\beta=1}$=76.22%) and 96.42% for intuitive judgments (macro-average $F_{\beta=1}$=67.27%).

Keywords: Natural Language Processing, document classification, clinical data, medical discharge summaries, obesity.

1. INTRODUCTION

Medical institutes usually store considerable amount of valuable information (patient data) as free text. Such information has a great potential in aiding research related to diseases or improving the quality of medical care. The size of document repositories makes automated processing in a cost-efficient and timely manner an increasingly important issue. The intelligent processing of clinical texts is the main goal of Natural Language Processing [1] for medical texts.

In this work, we introduce our system for identifying morbidities in the flow-text parts of clinical discharge summaries. The system was designed and implemented for the **Obesity Challenge** organized by the Informatics for Integrating Biology and the Bedside (I2B2), National Center for Biomedical Computing in spring 2008. The full paper with more detailed description is published as the online supplement of this study, and is available at www.jamia.org.

2. BACKGROUND

The importance of applying Natural Language Processing techniques to facilitate processes in clinical care and clinical research that require the analysis of textual data is clearly evidenced by the increasing number of publications and case studies related to the topic.

There were several shared tasks in the past few years that involved multi-label classification of clinical documents. The smoker challenge organized by I2B2 in 2006 [2] targeted the identification of the patient's smoker status. The clinical coding challenge [3] organized by the Computational Medicine Center of Cincinatti Children's Hospital in 2007 focused on the assignment of ICD codes to radiology reports to enable automated billing.

A. The obesity challenge

The target diseases of the Obesity Challenge included **obesity** and its 15 most frequent co-morbidities exhibited by patients, while the target labels corresponded to expert judgments based on ***textual evidence*** and ***intuition***. The development of systems that can successfully replicate the decisions made by obesity experts would be desirable to facilitate large scale research on obesity, one of the leading preventable causes of death [4-6].

For a more comprehensive description of the task and the data, see www.i2b2.org/NLP/ and [7].

B. Related work

Even though several results are reported in peer-reviewed literature on medical text classification (e.g. [8-10]), the most obvious references to work related to this study are the systems submitted to the same challenge by other participants.

The two main approaches of participants were the construction of rule-based dictionary lookup systems and statistical classifiers based on the Bag-of-Words (or bi- and trigram) representation of documents.

The dictionaries of rule-based systems mostly consisted of the names of the diseases, and their various spelling variants, abbreviations, etc. One team also used other related clinical named entities [11]. The dictionaries used were constructed mainly manually (either by domain experts [12] or computer scientists [13]), but one team applied fully automatic approach to construct their lexicons [14].

Machine learning methods applied by participating systems ranged from Maximum Entropy Classifiers [15] and Support Vector Machines [11] to Bayesian classifiers (Naïve Bayes [16] and Bayesian Network [17]). These systems showed competitive performance on the frequent classes but had major difficulties in predicting the less represented negative and uncertain information in the texts.

C. Our approach

Based on our previous experiences in similar tasks [18, 19] we observed that the classic word uni-, bi- or trigram (or in general n-gram) of words representation is not well suited to specific medical text classification problems like the obesity challenge, regardless of the learning method applied. This is mainly because the target pieces of information are in just a few sentences (possibly fragmented over the text) and the majority of the text is irrelevant to the problem.

In this sense the obesity challenge is more like an Information Extraction task, which gathers the relevant information from scattered sentences of the document, then makes the document-level decision based on the extracted information.

These aspects motivated us to develop a rule-based system to the challenge that exploits the lists of keywords that trigger important sentences (that is, the names and various spellings of the actual disease) and to implement a simple context analyser that enabled the correct prediction of negative and uncertain information in text. We applied statistical methods to complement, assist and speed-up manual work wherever it proved to be possible.

The system can be tested online at www.inf.u-szeged.hu/rgai/obesity. The most important resources of our system can be downloaded from the same site and are free for re-use if properly acknowledged.

3. METHOD

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.

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).

A. 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 occurences 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 judgment of relevance (information on the patient instead of family members, etc.) and an analysis of context to detect negation and 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 (the rules were executed in order, and once a rule was matched, the system assigned the relevant classification):

Classify a document as:

1. YES if any terms were matched in an assertive context
2. NO if any terms were matched in a negative context
3. QUESTIONABLE if any terms were matched in an uncertain context
4. UNMENTIONED if none of the previous steps triggered a different labeling.

B. Intuitive model

Our intuitive model was based on the textual model. That is, we attempted to discriminate the documents classified as UNMENTIONED by our textual classifier to intuitive YES or NO classes. When the textual system assigned a label that was different from UNMENTIONED, we accepted that decision as an intuitive judgment as well. Although somewhat simplistic,,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: names of associated drugs and medication, phrases related to certain social habits of the patients (e.g. cigarette for hypertension), tension values, weight, etc. While the phrases were collected using a semi-automated procedure similar to the one used to set up the disease term dictionaries, the numeric expressions describing relevant biomarkers were constructed by hand. Since these terms usually contained implicit information on the corresponding disease, it made no sense to evaluate their context for uncertainty. That is, the lists gathered specifically for the intuitive task 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** documents:
    a. intuitive **YES** if any intuitive-terms were matched in an assertive context
    b. intuitive **YES** if a numeric expression was below/above the predefined threshold
    c. classify a document as an intuitive **NO**.

C. System components

1) Keyword / Excluding term selection

The terms included in the dictionaries were gathered semi-automatically: we filtered them according to their frequency (infrequent terms were discarded in order to reduce the number of term-candidates and avoid overfitting on the data) and then ranked each term according to their positive class (**YES**) conditional probability scores (*p(yes|word)*). We evaluated the top ranked terms and added the meaningful ones to the corresponding disease-name dictionary manually. This way a 95% complete dictionary could be gathered quite rapidly – only the most frequent and reliable few dozens of keywords had to be evaluated manually for every disease.

Next, we collected pseudo terms (i.e. longer phrases containing a previously added disease term that are irrelevant to the disease) using a similar semi-automated procedure. This step was performed so as to avoid the overfitting of the dictionary lookup system (e.g. *'depression'*, but not *'st. depression'* or *'hypertension'* but not *'pulmonary hypertension'*).

The disease name dictionaries we collected were then extended with a few spelling variants manually, to handle different spellings of the same term.

2) Irrelevant contexts

We also made use of an UNMENTIONED dictionary that triggered the exclusion of the text from further processing. This way we excluded sections under headings like *'FAMILY HISTORY:'* and also phrases like *'son with…', 'family history of…'* from further processing. To define the scope of irrelevant phrases, we used the same context-identifier as that for negation and uncertainty detection (see below).

3) Negation / Uncertainty detection

The system with the above-described components was able to tag documents with YES labels or leave them as UNMENTIONED. Doing this, we also extracted sentences with disease names from YES-tagged QUESTIONABLE & NO documents and these sentences served as the basis for implementing a simple negation and uncertainty detection module. This exploited a list of negation / uncertainty cues and a list of delimiters (which triggered the end of scope). This approach is similar to NegEx [20]. Our biomedical text corpus annotated for negation and uncertainty [21] also demonstrates that this simple scope resolution approach works well for clinical texts.

4) Intuitive terms

We extended the system with ***intuitive dictionaries*** that triggered intuitive YES labels. These dictionaries were used to classify a document 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.

5) Biomarker expressions

We also added a model that looked for numeric expressions preceding or following certain keywords (that is, biomarker expressions) in the text to classify intuitive YES documents. 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*.

4. RESULTS

According to the official evaluation, our system achieved an F-macro score of 84% on the train for our best model (which degraded to 76% on the test set), and an intuitive F-macro score of 82% on the train set (which degraded to 67% on the test set) – detailed results can be seen in Tables 1-2 . This system came **sixth** in the textual F-macro ranking and **second** in the intuitive F-macro ranking (**third** best and **second** best micro-averaged scores, respectively). 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 judgment), and these classes had many more examples than the QUESTIONABLE and textual NO classes.

5. DISCUSSION

Our intuitive model was based on the textual model. This is why we got a worse performance in intuitive QUESTIONABLE tagging on the test data: we neglected textual UNMENTIONED documents that had an intuitive QUESTIONABLE label because there were too few of them in the training data to model this phenomenon, especially without background medical knowledge. Our system achieved the second best result on the previously unseen test set for both the micro- and macro-averaged evaluation (intuitive task). The good micro ranking tells us that the dictionaries we collected had a good coverage compared to other participants, while our second place in macro ranking confirms that predicting intuitive QUESTIONABLE cases also proved rather difficult (or even impossible) for the other participating systems as well.

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

We should add here that the main evaluation metric of the challenge was the macro-averaged F-measure. This metric gave special emphasis to the rare NO & QUESTIONABLE classes, which means that a few dozen examples had a major impact on the results.

This explains both the worse results on the test set (it was particularly hard to model these infrequent classes), and some seemingly strong drops (e.g. for *osteoarthritis*) or increases (e.g. for *obesity*) in performance for particular diseases. Micro averaged results, which take all document-label pair into account with a uniform weight, are more stable. Moreover, our third place in the micro ranking surely confirms that our disease term dictionaries had a reasonably good coverage (compared to other systems), while our context analyzer overlooked some NO & QUESTIONABLE cases (sixth place in macro ranking).

We suppose that the relatively good results achieved by our model are due to the high-precision term-dictionaries and context-analysis rules. We argue that such simple solutions are efficient whenever the classification depends on the presence or absence of certain single facts (assertions) in the text. In such problems, usually one sentence (in some cases, 2-3) contains the target information. This means that the information can be extracted using a simple approach based on dictionary lookup and modifier detection; and the recognition of complex dependencies in the document is not necessary.

For a detailed analysis and comparison of the submitted systems and their performance, see [7].

6. CONCLUSIONS

As regards the classification accuracy scores, the method proposed here looks quite promising for the automated processing of large datasets to gather information on obesity and related diseases. Classes with a few hundred training examples for each disease (**YES & UNMENTIONED** for textual and **YES & NO** for intuitive annotation) generally achieved a micro-averaged F-measure of around 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 most of the documents. We should mention here that the manual filtering of the synonym lists (which were collected using statistical methods) required no more than 10 minutes per disease on average, and the lists used for context-analysis seemed to be independent of the particular disease. The more time-consuming step was the manual evaluation of singleton documents that contributed to less than 1% of the system performance. These points make us think that our approach could be scaled up to a larger set of diseases without much effort.

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 probably lead to a substantial improvement in performance on these particular classes as well. Overall, we believe that our results demonstrate the feasibility of our approach for classifying clinical records and also show that even very simple systems with a shallow linguistic analysis can achieve remarkable accuracy scores for classifying clinical records on a limited set of concepts.

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; and the anonymous JAMIA reviewers for their valuable comments and suggestions. This work was supported in part by the NKTH grant of the Jedlik Ányos R&D Programme 2007 of the Hungarian government (codename TUDORKA7).

An earlier version of this paper [22] was presented at the NLP Challenge Workshop, sponsored by the i2b2 National Center for Biomedical Computing in November 2008. The workshop papers were not published.

References

1. Ananiadou S., McNaught J., 2005. Text Mining for Biology and Biomedicine. Artech House Publishers
2. Uzuner O., Goldstein I., Luo Y., and Kohane I., 2008. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 15(1):14–24.
3. Pestian J.P., Brew C., Matykiewicz P., Hovermale D.J., Johnson N., Cohen K. B., and Duch W., 2007. 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.
4. Allison D.B., Fontaine K.R., Manson J.E., Stevens J., VanItallie T.B., 1999. Annual deaths attributable to obesity in the United States. J. Am. Med. Assoc. 282 (16): 1530–8.
5. Mokdad A.H., Marks J.S., Stroup D.F., Gerberding J.L., 2004. Actual causes of death in the United States, 2000. J. Am. Med. Assoc. 291 (10): 1238–45.
6. Barness L.A., Opitz J.M., Gilbert-Barness E., 2007. Obesity: genetic, molecular, and environmental aspects. Am. J. Med. Genet. A 143A (24): 3016–34.
7. Uzuner O., 2009. Recognizing Obesity and Co-morbidities in Sparse Data. J Am Med Inform Assoc. 2009; **[typesetter please place issue and page numbers here, article is in current issue]**.
8. Wilcox A.B., Hripcsak G., 2004. The Role of Domain Knowledge in Automating Medical Text Report Classification. J. Am. Med. Inform. Assoc., 10(4): 330 - 338.
9. Hazlehurst B., Frost H.R., Sittig D.F., Stevens V.J., 2005. MediClass: A System for Detecting and Classifying Encounter-based Clinical Events in Any Electronic Medical Record. J. Am. Med. Inform. Assoc. 12(5): 517 - 529.

10. Pakhomov S.V.S., Buntrock J.D., Chute C.G., 2006. Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. J. Am. Med. Inform. Assoc. 13(5): 516 - 525.
11. Savova G, Clark C, Zheng J, Cohen KB, Murphy S, Wellner B, Harris D, Lazo M, Aberdeen J, Hu Q, Chute C, Hirschman L. 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.
12. Childs LC, Taylor RJ, Simonsen L, Heintzelman NH, Kowalski KM, Enelow R. Description of the Lockheed Martin / SAGE Analytica System for the i2b2 Challenge in Natural Language Processing for Clinical Data. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2008.
13. Solt I, Tikk D, Gál V, Kardkovács ZT. Context-Aware Rule Based Classifier for Semantic Classification of Diseases in Discharge Summaries. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2008.
14. Yang H, Spasic I, Keane JA, Nenadic G. Combining Lexical Profiling, Rules and Machine Learning for Disease Prediction from Hospital Discharge Summaries. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2008.
15. 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.
16. 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.

17. Matthews MP. Bayesian Networks and the i2b2 Obesity Challenge. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2008.
18. Szarvas Gy., Iván S., Bánhalmi A., Csirik J., Automatic Extraction of Semantic Content from Medical Discharge Records. WSEAS Transaction on Systems and Control 2006; 1(2): 312-317.
19. Farkas R, Szarvas Gy. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics 2008; 9(S3):S10.
20. Chapman W.W., Bridewell W., Hanbury P., Cooper G.F., and Buchanan B.G., 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 2001 5:301–310.
21. Vincze V., Szarvas Gy., Farkas R., Móra Gy., and Csirik J., 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008, 9(11): S9
22. Szarvas Gy, Farkas R, Almási A, Vincze V, Hegedűs I, Busa-Fekete R, Ormándi R. Simple Approaches to Disease Classification Based on Clinical Patient Records. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2008.

TABLES

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.