Research Paper ■

Semantic Classification of Diseases in Discharge Summaries Using a Context-aware Rule-based Classifier

ILLÉS SOLT, DOMONKOS TIKK, VIKTOR GÁL, ZSOLT T. KARDKOVÁCS

Abstract **Objective:** Automated and disease-specific classification of textual clinical discharge summaries is of great importance in human life science, as it helps physicians to make medical studies by providing statistically relevant data for analysis. This can be further facilitated if, at the labeling of discharge summaries, semantic labels are also extracted from text, such as whether a given disease is present, absent, questionable in a patient, or is unmentioned in the document. The authors present a classification technique that successfully solves the semantic classification task.

Design: The authors introduce a context-aware rule-based semantic classification technique for use on clinical discharge summaries. The classification is performed in subsequent steps. First, some misleading parts are removed from the text; then the text is partitioned into positive, negative, and uncertain context segments, then a sequence of binary classifiers is applied to assign the appropriate semantic labels.

Measurement: For evaluation the authors used the documents of the i2b2 Obesity Challenge and adopted its evaluation measures: F_1-macro and F_1-micro for measurements.

Results: On the two subtasks of the Obesity Challenge (textual and intuitive classification) the system performed very well, and achieved a F_1-macro = 0.80 for the textual and F_1-macro = 0.67 for the intuitive tasks, and obtained second place at the textual and first place at the intuitive subtasks of the challenge.

Conclusions: The authors show in the paper that a simple rule-based classifier can tackle the semantic classification task more successfully than machine learning techniques, if the training data are limited and some semantic labels are very sparse.

■ **J Am Med Inform Assoc.** 2009;16:580–584. DOI 10.1197/jamia.M3087.

Introduction

Biomedical text mining has become a thriving field because it proved its efficiency in a wide scope of application areas, such as the identification of biological entities in text,[1] assigning insurance codes to clinical records,[2] facilitating querying in biomedical databases,[3] etc. For a survey see Cohen and Hersh, 2005.[4] Discharge summaries offer a rich source of information for information extraction (IE) tasks, including classification. Several open challenges have been announced in this field: automated assignments of insurance codes to radiology reports[5] and smoking status identification task.[6]

The processing of textual medical records like discharge summaries facilitates medical studies by providing statistically relevant data for analysis. Analysis of a particular disease and its comorbidities on sets of patients is an example of this. The findings drawn from connections observed between elements of a set of diseases are of key importance in treatment and prevention issues.

In this paper we present results on the i2b2 Obesity Challenge shared task, which is a multiclass multilabel classification task focused on obesity and its 15 most common comorbidities (termed *diseases*). For each document, the task was to assign for each disease one of the following semantic labels: present, absent, questionable, or unmentioned (full description in Uzuner[7]).

The problem of the Obesity Challenge is an atypical, *two-dimensional* classification problem with disease and semantic dimensions.

The top 10 solutions are dominated by rule-based systems, while, interestingly, no machine learning based approach can be found among them (see the survey of Uzuner[7] and the online only version available at www.jamia.org).

Rule-based text classifiers (aka *expert systems*) were widespread (see, e.g., Hayes et al[8]) before the steady growth of

Affiliations of the authors: Department of Media Informatics and Telematics, Budapest University of Technology and Economics (IS, ZK), Budapest, Hungary; Institute of Computer Science, Humboldt University in Berlin (DT), Berlin, Germany; Department of Computer Science, Australian National University (VG), Acton, Australia.

The authors thank György Biró for his machine learning related tips, and Ferenc P. Szidarovszky for a part of text preprocessing, and finally Katalin Tóth, MD, for giving some domain-specific hints for rule creation. The authors thank the challenge organizers for providing them with this invaluable dataset and research experience; special thanks to Özlem Uzuner, for her role as organizer and correspondent.

Domonkos Tikk was supported by the Alexander von Humboldt Foundation.

Correspondence: Illés Solt, Department of Media Informatics and Telematics, Budapest University of Technology and Economics, 1117 Budapest, Magyar tudósok krt. 2, Hungary; e-mail: <illes.solt@tmit.bme.hu>.

Received for review: 12/02/08; accepted for publication: 04/07/09.

computational capacity made machine learning approaches more popular. The rule-based approach is often criticized due to the knowledge acquisition bottleneck (Sebastiani[9]). That is, each rule must be manually created, and the portability and flexibility of such systems are often very limited. These concerns are valid at categorization problems, where the rules are domain-dependent and the semantics of categories may shift. However, if the expert knowledge is available in knowledge bases (such as ontologies, typical for the biomedical domain), and the rules can be generated automatically, the overhead of manual processing can be minimized to error analysis. Consequently, expert and rule-based systems are often applied for different problems on medical domain (see, e.g., Zeng et al[10] and Chi et al[11]).

In the medical field, there is a growing need for interactive systems; however, the challenge did not address this aspect. Health experts usually do not trust a system that acts like a black-box, but instead they want to verify the evidences that support the decision made. Our system is transparent for humans, while a system that is using sophisticated and, hence, not easy-to-understand machine learning techniques may require additional efforts to achieve this goal.

Next we describe our context-aware rule-based classifier, present its performance on the i2b2 Obesity Challenge, and briefly discuss the results and lessons learnt from our study. For the community, we provide an online appendix to this paper (available as an online data supplement at www.jamia.org) and on-line demo (available at categorizer.tmit.bme.hu/~illes/i2b2/obesity_demo).

Methods

Problem Definition

In general, the problem is a multiclass multilabel classification task, but the applied semantic labeling and the selected evaluation criteria make the problem setting unusual. For each disease the annotators labeled documents with Y, N, Q, and U. Here Y means that the disease is *present*, N means that it is *absent*, Q means that it is *questionable* in the patient, and U means that the disease is *unmentioned* in the document. The labeling was performed in two different aspects: textual and intuitive. Textual contains only judgments strictly based on text; intuitive is based also on implicit information found in narrative text (and here only Y, N, and Q labels are assigned).

Systems were evaluated primarily based on their averaged F_1-macro value and secondarily based on their averaged F_1-micro value calculated over the 16 disease classes.

The problem can be considered as a two-dimensional task, where documents should be assigned to 16 disease classes and 4 (resp. 3) semantic classes. The two sets of classes are not independent.

- The semantic classes are distributed very unevenly: there exist classes even without training examples (6 textual N, 7 textual Q, and 5 intuitive Q classes out of 16).
- For the textual task, the annotators had to assign labels based on text fragments of documents, however, these fragments are not marked in the training set. The existence of such mark up could be extremely helpful when developing an automatic classifier.

Preprocessing of Text

To address the special characteristics of the tasks and assist manual exploration of the data, we first applied the following preprocessing steps:

Abbreviation Resolution. Resolving abbreviations aids non-professionals in understanding medical texts, by contrast decoding and disambiguating abbreviations in medical texts can improve the accuracy of information extraction.[12] It can be observed that periods, spaces and letter capitalization are used almost freely. In addition to dictionary lookups and web searches, we developed a regular expression driven string replacement dictionary to decode all occurrences of relevant abbreviations (see the online only version).

The style of abbreviations in a record may be characteristic of the health care institution where it originates. The deidentification of medical records usually covers only the obfuscation of named entities, however, we found that documents could be clustered by their abbreviation style. Consequently, we suspect that an adversary could bind documents to the note taking person observing the clusters and corresponding patterns (with better than uniform probability), however, we cannot verify this claim due to the lack of data. Nevertheless, we suggest that concealment of abbreviation styles should be the part of the deidentification process.

Identify Discharge Summary Elements (Zones). The exploration of documents is facilitated by a visually enhanced version with text being chunked into zones. Zones were recognized by the surface features of the headings (e.g., *Clinical Course*, *Diagnosis*, etc).

Classifiers

Next we describe our context-aware rule-based classifier for both the textual and intuitive subtasks. The classifiers only differ in the set of rules. When creating rules, we intended to mimic the work of the annotators. We found most documents to be labeled with Y/N/Q (textual) if at least one *mention* of the disease occurred explicitly in the text, and with U otherwise. A mention can be various forms of the disease name or a directly associated term/phrase. Annotators had to decide on the label based on the *context of the mentions*.

The Baseline Y-Classifier

We built a dictionary that included for each disease a set of *clue terms*. The sets contain the disease's name and its alternatives:

- Abbreviations (e.g., *HTN* for *hypertension*),
- Synonyms (e.g., *cardiac failure* for *heart failure*),
- Plain English equivalents (e.g., *high cholesterol*),
- Spelling variants (e.g., *gall stone* instead of *gallstones*),
- Frequent typos (e.g., *dislipidemia* instead of *dyslipidemia*),
- Suffixed forms (e.g., *gouty*),
- Related terms (e.g., *dyslipidemia* for *Hypercholesterolemia*).

Based on the set of clue terms, a rule-based *baseline classifier* was created for Y label, which assigned the label if any of the clue terms was present in the document.

Family History

Obviously, the baseline classifier fails to distinguish between mentions occurring in different zones. However, the zone information can be crucial. A typical case is the group of *family*

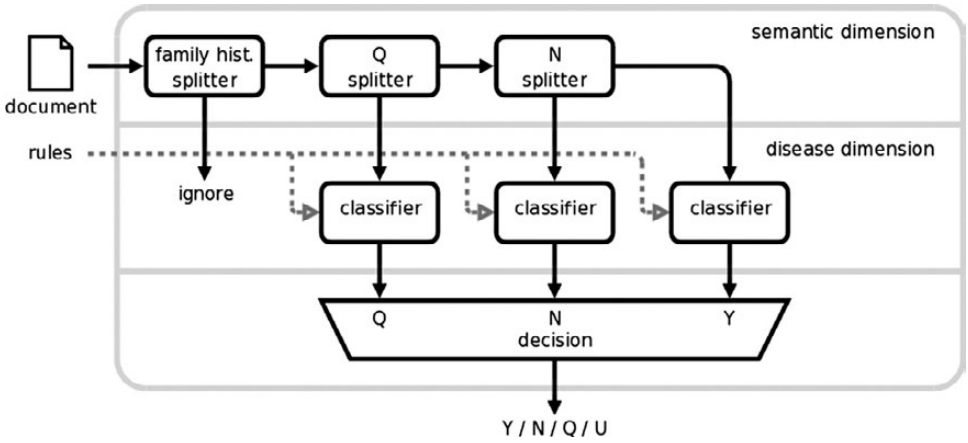

Figure 1. Document processing and classification workflow.

history zones, in which the text is not directly related to the patient's health. Therefore, we removed full zones and other fragments containing family history. During development, the removed fragments were systematically checked for false-positive matches. We ignored text fragments with family history and allergies (see details in the online only version).

Context-aware Partitioning, N- and Q-Classifiers

The baseline classifier was a binary classifier that only assigns Y labels. We found documents labeled with N (or Q) if the disease was mentioned in a negative (or uncertain) context. Since the primary evaluation function is F_1-macro, correctly assigning the rare N and Q labels to documents was a key issue.

First, we addressed the *semantic dimension* of the task. All documents were partitioned into family history, negative, uncertain, and positive contextual segments by a sequence of *splitters*, which reduced the 4-way partitioning problem to the binary case.

Inspired by NegEx,[13] a framework to identify negative statements in medical text, we used trigger phrases to recognize negative and uncertain contexts. The differences between our solution and NegEx are the following. We use a different set of triggers and scope-terminating words. We built our solution for the Obesity Challenge using solely the given training records, surprisingly, without resulting in overfitting to the medical domain (see Section 4.). As an improvement over NegEx, we also applied midscope and word prefix triggers. We provide some illustrative fragments with triggers (bold) not included in the latest version of NegEx:

1. Pre- and postscope triggers
   Example: "[?history of diabetes.]$_{\text{uncertain}}$" yields Q for *Diabetes*, "[outpatient *OSA* **screen**.]$_{\text{uncertain}}$" yields Q for *OSA*
2. Scope limiting triggers
   Example: "[**No** other significant past medical history]$_{\text{negative}}$ **besides** *hypertension*." yields Y for *Hypertension*
3. Mid-scope triggers
   Syntax: [words] trigger [words]
   Example: "[Right lower lobe pneumonia **versus** *CHF*.]$_{\text{uncertain}}$" yields Q for *CHF*
4. Word-prefix triggers
   Syntax: trigger (non-space characters) Example: "Abdomen: soft, [non-obese]$_{\text{negative}}$, and [non-tender]$_{\text{negative}}$" yields N for *Obesity*

Semantic classification is followed by tackling the *disease dimension*. Once the documents were partitioned based on the context, they were fed to the appropriate binary baseline classifiers to obtain Q, N, and Y judgments. Thus, more than one label could be assigned to a document. The conflicts were resolved as follows. The Q-classifier is the least certain one: if a disease appears in both an uncertain and positive/negative context, then the latter rules out the former. Similarly, any positive mention takes precedence over negative mentions. Therefore, we applied the 3 classifiers in a cascade like pipeline: first Q labels were assigned, then N labels, finally Y labels (labels are overwritten), and unlabeled documents received U labels. The document processing and classification workflow is depicted in Fig 1.

Our solution separates semantic and disease dimensions: the Q-, N- and Y-classifiers are based on the same set of rules that makes the classification simple, but more importantly, it allows us to use one common training set (per disease) for all classifiers, which is crucial due to the very limited number of Q and N labels.

Intuitive Judgments

At this point we have textual judgments. The text- and rule-based approach of the baseline classifier was less powerful on intuitive judgments, while the corpus had virtually no textual indicators recognizable by human experts or rule-based learners. Still, we found some simple rules that, in addition to those created for textual judgments, improved performance on the intuitive set. These rules include:

- Disease-specific, non-preventive medications and their brand names (e.g., *SSRI* or *Zoloft* for *Depression*)
- Related procedures (e.g., *panniculectomy* for *Obesity*)
- Symptoms with very high correlation (e.g., *leg ulcers* for *Venous Insufficiency*)

Example

Let us illustrate the complete work of the text splitter and classifier method, on a paragraph taken from the test document #1,058:

"CARDIAC RISK FACTORS: *Hypertension*, smoking [**family history of** *coronary artery disease*]$_{\text{family}}$ and CVA. [**No** *diabetes*.]$_{\text{negative}}$ she is [**not** postmenopausal]$_{\text{negative}}$ [**no** history of *elevated cholesterol*.]$_{\text{negative}}$ [**No** previous myocardial infarction history.]$_{\text{negative}}$"

Table 1 ■ F_1-Macro and F_1-Micro Results of Our Best Submissions

	Textual		Intuitive	
	F_1-Macro	F_1-Micro	F_1-Macro	F_1-Micro
Disease				
Asthma	0.9434	0.9921	0.9784	0.9894
CAD	0.8561	0.9256	0.6122	0.9192
CHF	0.7939	0.9355	0.6236	0.9315
Depression	0.9716	0.9842	0.9346	0.9539
Diabetes	0.9032	0.9761	0.9682	0.9729
Gallstones	0.8141	0.9822	0.9729	0.9857
GERD	0.4880	0.9881	0.5768	0.9131
Gout	0.9733	0.9881	0.9771	0.9900
Hypercholesterolemia	0.7922	0.9721	0.9053	0.9072
Hypertension	0.8378	0.9621	0.8851	0.9283
Hypertriglyceridemia	0.9732	0.9980	0.7981	0.9712
OA	0.9594	0.9761	0.6286	0.9589
Obesity	0.4879	0.9675	0.9724	0.9732
OSA	0.8781	0.9920	0.8805	0.9939
PVD	0.9682	0.9862	0.6348	0.9763
Venous insufficiency	0.8403	0.9822	0.8083	0.9625
Overall	0.8000	0.9756	0.6745	0.9590

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

For this paragraph the overall classifier system would assign Y for *hypertension*, N for *Diabetes* and *Hypercholesterolemia*, and U for all others (note: including *CAD*).

Results

Table 1 includes the main evaluation measures of our best submissions for both subtasks as calculated by the organizers. We also present the confusion matrices in Table 2a and 2b.

The choice of F_1-macro for the main evaluation metric motivated participants to focus on sparse classes. Applying F_1 also implies that misclassification from one label to another is penalized in a degree dependent on the distribution of labels, and not according to the degree of the mistake's seriousness. For example, in the textual case, misclassifying a document as N instead of its gold standard Y, is penalized less than misclassifying it as Q, though the semantics of the labels suggest the opposite. Not concentrating on the sparsest Q labels—the approach followed by many other participants—would give an upper bound on F_1-macro of three-fourth (textual), and two-thirds (intuitive). Our system was able to break both upper bounds.

We achieved a second position in textual and a first position in the intuitive subtask with our system that was built from scratch: it does not incorporate any third party software, or other external resources. In comparison with the other top teams' submissions, the main advantage of our system over other participants' was the low number of misclassified Q labels, in other aspects its performance is on par with those ones.

Table 2a ■ Textual Confusion Matrix on the Test Set

	Y	*N*	U	Q
Y	2117	8	66	1
N	10	41	14	0
U	71	16	5681	2
Q	5	0	3	9

Rows represent the gold standard, while columns indicate the label assigned by our system.

Table 2b ■ Intuitive Confusion Matrix on the Test Set

	Y	N	Q
Y	2106	178	1
N	110	4989	1
Q	6	7	1

Rows represent the gold standard, while columns indicate the label assigned by our system.

Our system is scalable: the number of rules grows linearly with the number of classes (health conditions), while the context identification components (text splitters) can be entirely reused without modification for the medical domain, thus reducing development time. Classification rules do not have to be created manually, they can also be bootstrapped from domain-specific ontologies (like UMLS[14]) or obtained from machine learners.

Though only the most common comorbidities of obesity were selected for this challenge, the training examples were still very limited in number, forecasting even sparser data when scaling up to other conditions. Our system performs relatively stable across all classes: F_1-macro is on average higher on classes with more data, but drops only slightly for sparser classes (see Fig 2a in online only version).

We also investigated the portability of our system to other domains. For comparison, we evaluated NegEx and our system on the biomedical corpus provided by the BioNLP2009 Shared Task[15] that featured negation and speculation annotations. The results summarized in Table 3 show that our context-aware system slightly outperforms NegEx also on the biomedical domain, without any corpus-dependent adaptation.

Conclusions

In this paper we reported on our approach for the i2b2 Obesity Challenge. We developed a context-aware rule-based classification model that was able to cope successfully with this unusual multiclass multilabel classification task by a two-dimensional approach that handles disease and se-

Table 3 ■ Performance of Assertion Classification on the BioNLP2009 Corpus

Semantic Class	System	Precision	Recall	F_1-Measure
Negation	NegEx	0.4070	0.5152	0.4548
Negation	Our original system	0.5015	0.4626	0.4813
Negation	Our adapted system	0.7662	0.7460	0.7561
Speculation	NegEx	0.0000	0.0000	0.0000
Speculation	Our original system	0.1340	0.0236	0.0402
Speculation	Our adapted system	0.6049	0.5472	0.5746

We compared NegEx (v1.01 of Imre Solti's implementation, code.google.com/p/negex/), our system developed for the i2b2 Obesity Challenge, and the adaptation of our system towards the BioNLP2009 Shared Task (for comparison only).

mantic classes separately. Thus, we could exploit efficiently the very limited training data.

Using manually fine-tuned regular expressions for identifying contexts and diseases, the performance could be slightly increased compared to those systems using preexisting domain specific tools and resources. By keeping the system's structure simple, its output can be presented in a way that makes human verification quick and easy; our system is thus suitable for both fully automated black-box operation and incorporation into interactive systems. Our system scales up well with the number of classes and is portable to other not necessarily clinical corpora. As an extension of this paper, we also provide a detailed online appendix that contains the resources we applied and an on-line demo with most of the functionalities of our system.

References ■

1. Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 2005;10(6):439–45.
2. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinform 2008;9(S3):S10.
3. Plake C, Schiemann T, Pankalla M, et al. PubMed as a graph. Bioinformatics 2006;22(19):2444–5.
4. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005;6(1):57–71.
5. International Challenge: Classifying Clinical Free Text Using Natural Language Processing. Available at: http://www.computationalmedicine.org/challenge/index.php. Accessed May 2009.
6. Uzuner Ö, Szolovits P, Kohane I. i2b2 workshop on natural language processing challenges for clinical records. Proc. of the Fall Symp. of the Amer. Med. Info. Assoc. (AMIA 2006), Washington, DC, 2006.
7. Uzuner Ö. Recognizing obesity and Co-morbidities in sparse data. J Am Med Inform Assoc 2009:
8. Hayes PJ, Andersen PM, Nirenburg IB, Schmandt LM. Tcs: A shell for content-based text categorization. In: 6th Conference on Artificial Intelligence Applications, 1990.
9. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002;34(1):1–47.
10. Zeng Q, Goryachev S, Weiss S, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6(1):30–8.
11. Chi CL, Street WN, Ward MM. Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. J Biomed Inform 2008;41(2):371–86.
12. Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in MEDLINE. Bioinformatics 2005;21(18):3658–64.
13. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:301–10.
14. Lindberg DA, Humphreys BL, McCray AT. The Unified medical language system. Methods Inf Med 1993;32:281–91.
15. Tsujii Laboratory [homepage on the Internet]. Tokyo, Japan: BioNLP'09 Shared Task on Event Extraction: in conjunction with BioNLP, a NAACL-HLT 2009 workshop, June 4-5 2009, Boulder, Colorado. Available from: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/. Accessed May 2009.