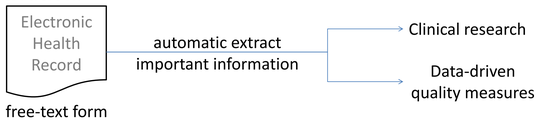
Automated Annotation on Colonoscopy Reports

Phuong Pham[1], Janyce Wiebe, PhD[1], Rebecca Hwa, PhD[1], Wendy W Chapman, PhD[2]

[1]University of Pittsburgh, Pittsburgh, PA; [2]University of California, San Diego, CA

Introduction

Motivation:



Objective:

- We laid out an annotation framework and evaluated it in a case study where quality measures were extracted from colonoscopy reports.
- In this case study, we chose a configuration which can be applied to multiple domains without medical and natural language processing (NLP) expertise but still has reasonable results.

#1: Annotation task

Identifying the problem we are working on (usually done by domain experts/researchers)

A formal representation is needed in this step for both applying automatic methods and using human annotators.

For example

> *Record yes if physician stopped the procedure early due to too much stool, patient discomfort,* ***etc.***

proc-aborted definition

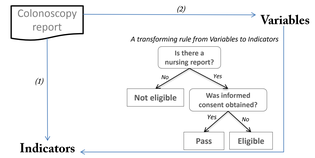
•**Automatic methods**: it is difficult to implement the case *etc.* in a system

•**Human annotators**: annotators will be confused about the case *etc.*

This step is domain-independent.

#2: Objective function

Identifying the information we want to annotate. We have 2 possible objective functions in this project



Choosing

1.Indicators: did not give good results. Indicators are more abstract (than Variables) and thus difficult to automatically predict given the simple annotating features used in this project.

2.Variables: gave interesting results, which are reported in *Table 2*. After extracted values of Variables, we can easily convert them into Indicators by deterministic rules.

Single concept variables are better choices for our framework configuration.

#3: Annotating features

Annotating features are representations of input data. We tried **bag of words** features

1. They are easy to apply to *new domains* (which is our framework configuration)
2. They can be *applied quickly* to evaluate different approaches
3. They empirically give *reasonable results*

#4: Annotating methods

We tried different kinds of method:

1. Simple method: **SimpleSQ** which is a text searching system
2. Rule-based method: **C-QUAL** which is proposed by Harkema et al (2011)[1]. This system is considered as the baseline in this project
3. Generic machine learning (ML) methods: Support vector machine (**SVM**) and Decision tree (**J48**). These methods are generic and domain-independent.

As shown in *Table 2*, the simple method worked well in many cases. Meanwhile, ML methods improved some cases where the simple method did not work well.

#5: Skewed data

Data in clinical domains is often skewed. We tried 2 popular re-sampling methods, which are generic and domain-independent, to handle skewed data (shown in *Table 1*), i.e. **over-sampling** and **down-sampling**.

As shown in *Figure 1* and *Figure 2*, re-sampling methods only helped minority classification and degraded majority class classification. Moreover, re-sampling methods only help when the data distribution is not balanced or extremely skewed. In our experiments, down-sampling outperformed over-sampling.

#6: Performance measures

We explored popular performance measures to evaluate different systems, i.e. *precision*, *recall*, *F-measure*, and *accuracy*. We looked at different views on these performance measures of our binary variables

Yes class performance:

Predict as ->	No	Yes
No	*TN*	*FP*
Yes	*FN*	*TP*

No class performance:

Predict as ->	No	Yes
No	*TP*	*FN*
Yes	*FP*	*TN*

- **Yes class performance**
- **No class performance**
- **Average class performance**: average of Yes class and No class performance
- **Majority class performance, Minority class performance:** if a variable has many *Yes*es, then the Majority class performance is the Yes class performance and vice versa

This step is domain-independent.

Conclusion

In our case study, it is possible to configure the framework to have a system which requires no medical or NLP expertise, and has reasonable results. All the experimented features and methods are generic and feasible to apply to different clinical domains. In this study, we showed that the configuration requires:

•Variables (objective function) should be single concept

•Run simple keyword searching first and accept variables whose performance is above a certain threshold.

•For other variables, we try learning machine methods. Re-sampling methods only help minority class classification when the variable is not balanced or extremely skewed.

[1] ***Harkema H**, Chapman WW, Saul M, et al. Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc 2011;18(Suppl 1): i150-6.*

Variable	Yes	No	Yes %
any-adenoma	128	325	28.26%
appendiceal-orifice	288	165	63.58%
ASA	407	46	89.85%
Biopsy	75	378	16.56%
Cecum	445	8	98.23%
ileo-cecal-valve	306	147	67.55%
indication-type	22	431	4.86%
informed-consent	289	164	63.80%
nursing-report	27	426	5.96%
prep-adequate-YES	225	228	49.67%
prep-adequate-NO	38	415	8.39%
prep-adequate-NOTMENTIONED	190	265	41.76%
proc-aborted	6	447	1.32%
withdraw-time	450	3	99.34%

Table 1: variable distribution

Variable	C-QUAL			SimpleSQ			J48-no sampling			J48-down sampling		
	P	R	F	P	R	F	P	R	F	P	R	F
any-adenoma	95.4	97.7	96.5	95.4	97.7	96.5	95.3	94.9	95.1	99.6	94.1	96.8*
appendiceal-orifice	98.1	87.2	92.3	96.8	83.0	89.4	91.1	93.4	92.2	79.2	90.2	84.3*
ASA	98.5	99.5	99.0	100	99.5	99.8	99.9	99.5	99.6	93.1	89.6	91.2*
Biopsy	42.8	98.7	59.7	42.5	60.0	49.7	70.7	61.8	65.9	89.4	80.2	84.5*
Cecum	98.9	98.7	98.8	98.8	89.9	94.1	98.2	99.4	98.8	NaN	0.0	NaN
ileo-cecal-valve	98.6	89.5	93.8	98.0	94.8	96.4	96.0	96.5	96.2	79.3	93.7	85.8*
indication-type	85.7	81.8	83.7	73.3	100	84.6	76.7	77.7	77.1	25.1	55.3	34.1*
informed-consent	99.2	81.0	89.1	97.0	88.2	92.4	98.2	98.6	98.4	88.9	94.7	91.7*
nursing-report	88.9	59.3	71.1	85.7	22.2	35.3	78.5	84.8	81.5	34.3	72.4	46.2*
prep-adequate-NO	57.1	21.1	30.8	92.9	34.2	50.0	60.6	44.8	51.2	38.6	47.3	41.6*
prep-adequate-NOTMENTIONED	80.9	95.8	87.7	69.3	82.1	75.2	86.6	87.8	87.2	86.9	90.5	88.7*
prep-adequate-YES	95.3	90.7	92.9	78.0	74.2	76.1	84.2	85.6	84.9	9.2	1.2	2.1*
proc-aborted	0.0	0.0	NaN	2.0	33.3	3.9	NaN	0.0	NaN	NaN	0.0	NaN
withdraw-time	100	99.8	99.9	100	100	100	99.5	99.8	99.6	NaN	0.0	NaN

Table 2: Performance of different systems. Because of the space limit, we do not report the result of SVM and over-sampling here, but the same trend with J48 was observed. * means the re-sampling method is significantly better/worse than the no-sampling method

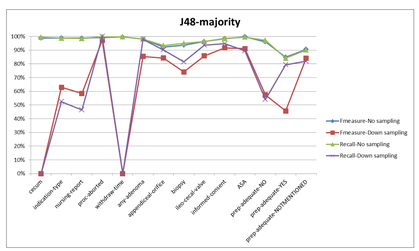


Figure 1: majority class F-measure and Recall of J48

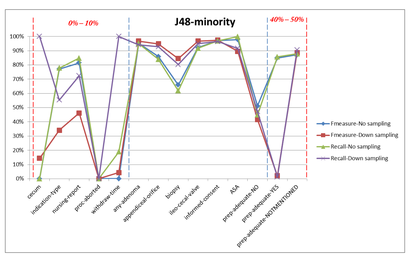


Figure 2: minority class F-measure and Recall of J48

This research is funded by the National Library of Medicine under grant 5R01LM010964.