Automatic assignment of diagnosis to medical reports

Adrián Lachata | Advisor: Jiří Hana | Faculty of Mathematics and Physics, Charles University in Prague, 2014

Question

Can a machine assign correct diagnosis code to a medical report?

Possible Usage

Secondary diagnosis | Backward checking | IntelliSense | Rare diagnoses

Thesis Overview

We developed a system for textual classification based on techniques of machine learning and natural language processing (NLP). Our system preprocesses a text, generates key features, filters them to make their number manageable, and finally applies a supervised classification algorithm. We used it for automatic assignment of diagnosis codes (using the ICD-10 classification) to **Czech textual medical reports**.

As a pilot, we have selected 5 diagnoses, focusing mostly on the I10 diagnosis (Essential hypertension). For these 5 diagnoses, the system has been trained and tested on a **corpus of one million medical reports**. The most promising results are for I10. For I10, we have also compared our automatic results with classification performed manually by doctors of medicine.

Why I10?

- The most frequent diagnosis
- Well understood by public

Chosen diagnoses

ICD-10	Description	Train	Test
I10	Essential (primary) hypertension (high blood pressure)	8.9	10.7
H66.0	Abscess externa	4.0	0.9
J00	Acute nasopharingitis (common cold)	3.9	2.6
K30	Dyspepsia	1.1	1.1
Z00.1	Routine health child examination	4.2	4.9

Table 1: Distribution of chosen diagnoses and their proportion (in %) in train data of 100 000 records and test data of 500 000.

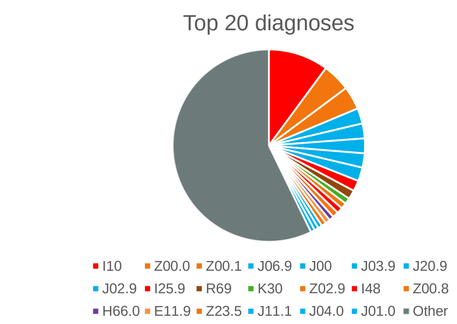

Figure 1: Distribution of top 20 diagnoses in training and testing data. The same colors represents diagnoses in same category (first letter).

Domain

Sample 1: Subjektivní: Quick 1,64 INR - Warf 3mg 1-0-0 5x týdně, a 2 x týdně 1,5-0-0. Objektivní: KP komp. AS zdá se pravidelná.

Sample 2: Závěr: Ukončena PN k datu 23.11.2010.

Sample 3: Subjektivní: Cítí sedobře

Sample 4: Subjektivní: Průjem jako její přítel. Špatně spí. Objektivní: KP komp. Závěr: Kontrola p.p., při průjmu. Dieta nutná. Poučena o rizicích Hypnogenu. Neukázněný pacient.

Sample 5: Objektivní: 6 250gr.Srdce,plice bpn.Vidí,slyší.Pevně drží hlavič-ku.Nutrilon prem.

Sample 6: Subjektivní: Bez akut. stesků. Objektivní: TK 160/95, KP komp. Terapie: Prestance 10/5 1-0-0, Flexove

Sample 7: Subjektivní: Cítí se dobře Objektivní: TK 130/80

Sample 8: Subjektivní: Měla virosu. Občas jí bolí za krkem. Závěr: Zkusí vyměnit polštář.

Stopwords generated by Inverted Document Frequency (**IDF**): objektivní, subjektivní, tknout, na, lék, v, a, pro, p, bez, -, a, k, odběr, být, závěr, nález, kontrola, k, mít, pravit, ad.

Our system

Figure 2: Process flow of the system we developed.

Impact of Preprocessing

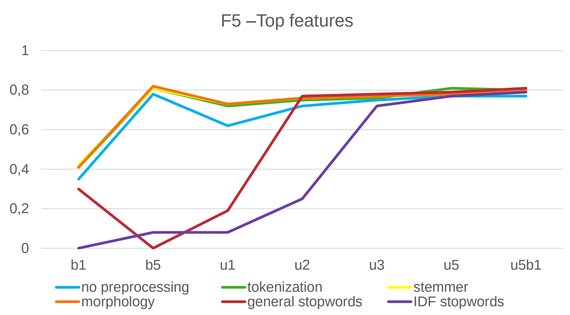

Figure 3: Impact of preprocessing for diagnosis I10 and Naive Bayes classifier. *b* means bigrams, *u* unigram and the number means hundreds. E.g. *u5b1* means 500 unigrams and 100 bigrams.

Impact of Feature Count

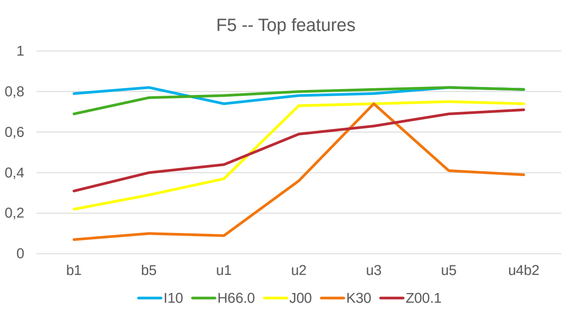

Figure 4: Impact of features count for diagnosis I10 and Naive Bayes classifier.

Impact of Classifiers

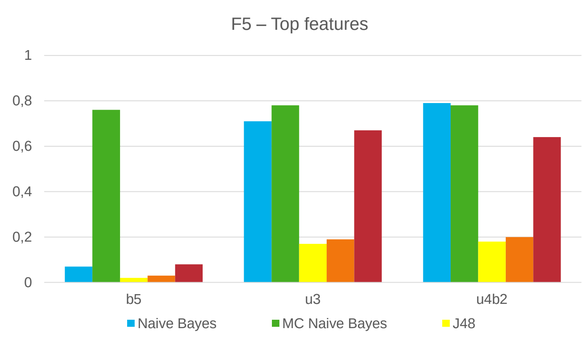

Figure 5: Impact of classification algorithms on diagnosis I10. MC means Meta Cost with given classifier and Cost Matrix set [0;1,5,0] in favor of Recall.

Note: Meta Cost with J48 was too computationally hard.

Impact of Features Filter

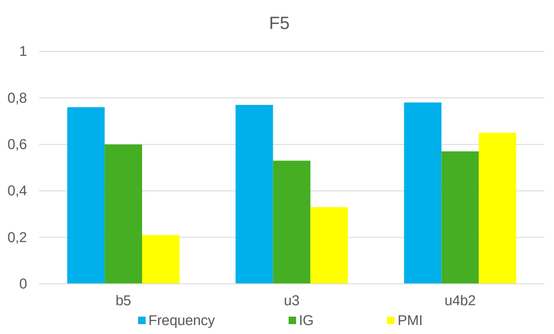

Figure 6: Impact of features filter algorithms for diagnosis I10 with classifier Naive Bayes in favor of Recall.

I10 Selected Features

IG unigrams			PMI unigrams			Frequency unigrams			IG bigrams			PMI bigrams			Frequency bigrams		
Feature	FF	Dg	Feature	FF	Dg	Feature	FF	Dg	Feature	FF	Dg	Feature	FF	Dg	Feature	FF	Dg
tknout	9 435	4 321	levotyp	10	10	zlepšit	2 460	66	objektivní tknout	3 672	2 154	lék potíž	13	13	zákonný zástupce	2 095	36
ia	413	325	ifirmacombi	8	8	bolestivý	2 448	28	pro lék	1 737	1 184	- 150/90	10	10	ošetření souhlasit	1 991	17
doctor's name	233	189	telmisartan	6	6	chtít	2 446	155	tknout -	343	292	- 135/70	10	10	vrozený vada	1 862	0
preskribce	370	245	lusopr	6	6	palpa	2 414	75	subjektivní pl	1 115	615	ia objektivní	10	10	štítný žláza	1 402	1
140/80	685	373	irbesartan	5	5	užívat	2 413	158	subjektivní pro	434	324	ebrantil 30	10	10	srdeční pravidelný	1 075	26

Table 2: Top unigrams and bigrams for diagnosis I10 selected by Information Gain (IG), Pointwise Mutual Information (PMI) and Frequency without IDF stopwords. All filters used morphology. FF – feature frequency, Dg – frequency of feature and diagnosis together.

Are the most frequent features the best?

Filter	TN	FP	FN	TP	P	R	F1	F5	ROC	K
Freq	37 307	53 781	346	8 600	0.14	0.96	**0.24**	0.78	0.77	**0.10**
IG	86 799	4 289	3 788	5 068	0.54	0.57	**0.55**	0.57	0.83	**0.51**
PMI	85 348	5 740	4 115	4 831	0.46	0.54	**0.50**	0.54	0.76	**0.44**

Table 3: Evaluation metrics for diagnosis I10 with 400 unigrams, 200 bigrams and classifier Naive Bayes in favor of Recall. TN – True Negative, FP – False Positive, FN – False Negative, TP – True Positive, P – Precision, R – Recall, F1 – F – Measure, ROC – arena under ROC curve, K – Kappa.

Machine vs Human

Filter	TN	FP	FN	TP	P	R	F1	F5	ROC	K
Doctor 1	46	6	19	29	0.83	0.60	0.70	0.61	---	0.50
Doctor 2	50	2	32	16	0.89	0.33	0.48	0.34	---	0.32
Doctor 3	49	2	29	20	0.91	0.41	0.56	0.42	---	0.38
IG	45	4	32	18	0.82	0.36	0.50	0.37	0.69	0.26
Freq	17	32	6	44	0.52	0.88	0.70	0.86	0.68	0.22

Table 4: Evaluation metric for three medical doctors and 400 unigrams with 200 bigrams used morphology and features filtered by frequency and Information Gain (IG). For legend see table 3.

All diagnoses

Dg	TN	FP	FN	TP	P	R	F1	F5	ROC	K
I10	439 541	30 968	23 435	26 862	0.46	0.50	0.48	0.50	0.78	0.42
H66.0	515 554	3 922	1 130	3 210	0.45	0.74	0.56	0.72	0.95	0.56
J00	503 877	6 710	9 823	3 396	0.34	0.28	0.29	0.26	0.77	0.28
K30	514 806	3 696	4 587	717	0.16	0.14	0.15	0.14	0.64	0.14
Z00.1	501 197	2 407	16 057	4145	0.6	0.21	0.31	0.21	0.64	0.12

Table 5: Evaluation metrics for all chosen diagnoses on testing data. 400 unigrams with 200 bigrams used morphology and features filtered Information Gain (IG). For legend see table 2.

Summary

Textual preprocessing

- Not necessary but can improve performance
- Removing non-alphanumerical characters
- Preserving blood pressure measurement (e.g.135/75) and drugs dosage (e.g. 1-0-1)

Features – 400 unigrams with 200 bigrams

- More than 300 unigrams did not make results significantly better
- Bigrams can reveal hidden relations
- 600 features were optimal for our computer performance

Features Filter – Information Gain

- The most frequent features even without stopwords are always the same for every diagnosis
- PMI chooses the most unique features for given diagnosis. The features can be data specific
- IG takes data and feature entropy into the account

Classifiers – Meta Cost with Naïve Bayes

- Naive Bayes has the best results with the least computation time
- Cost of not assigning was 5 times bigger than wrong assignment
- Meta Cost does not work worse than pure Naive Bayes

Evaluation – F5 and Kappa

- Lot of reports with content irrelevant to any diagnosis (good F5 by features selected by frequency)

Diagnosis I10

- Features selected by IG are highly relevant
- Bigrams show relation between I10 and drugs prescription
- When our system assigned I10 to a report, at least one doctor assigned I10 too

Machine vs Human

- Doctors had good Precision
- Only one doctor was aware of hidden association between I10 and drugs prescription
- Machine assigned I10 to the reports only about drugs prescription
- In all cases when the diagnosis was assigned incorrectly, there were also assigned incorrectly at least by one doctor