Representation and classification techniques for clinical data focused on obesity and its co-morbidities

Oana Frunza, MSc, Diana Inkpen, PhD
University of Ottawa, ON, Canada

Abstract

In this paper, we describe the three representation and classification techniques used in our submission for the I2B2 Shared-Task – Challenges in Natural Language Processing and Clinical Data. Our methods combine a simple bag-of-words representation, a UMLS (Unified Medical Language System) concept representation, and a noun-phrase representation with the AdaBoost classification algorithm. The results obtained for the two tracks of the competition, the textual and the intuitive track, show that the bag-of-words representation obtained better results than the other two.

Introduction

The I2B2 Shared-Task, Challenges in Natural Language Processing for Clinical Data is focused on obesity and its co-morbidities. The data released for the competition consists of discharge summaries of patients annotated with obesity and related diseases.

Each patient record went under a process of de-identification. The procedure consisted in an automatic and two manual passes for every record. The disagreements were solved using a third manual annotator. The doctor and patient names were replaced by random names extracted from the US Census Bureau[1] names dictionary. The phone numbers, ID numbers, and ages were replaced by randomly drawn digits for each of the numbers. The dates were replaced by surrogate ones. The new locations and hospital names were obtained by permuting syllables of the original ones. Valid cities, state names, and zip codes were added to the new generated locations.

The records released as data for the competition were document-level annotated with obesity information and its co-morbidities. Each document is a longitudinal record of a patient; the information that it contains is collected over time.

The goal of the challenge is the development of systems that accurately identify obese patients and the co-morbidities correlated with the disease.

Task Description

The task was focused on a data set of clinical patient records obtained as a result of a query-based retrieval on the stem "obes" extracted from the Research Patient Data Repository (RPDR).

Two types of annotation were supplied for each record: textual annotation – judgments made only on texts; and intuitive annotation – judgments made on implicit information in the narrative text (*e.g.,* computing the BMI (Body Mass Index) is considered part of an intuitive annotation process). Two doctors annotated in parallel the data, and the disagreements were resolved by the help of a third expert.

For the textual judgments the possible classes are: "Y" that stands for "Yes, the patient has the co-morbidity", "N" that stands for "No, the patient does not have the co-morbidity", "Q" that stands for "Questionable whether the patient has the co-morbidity", "U" that stands for "the co-morbidity is not mentioned in the record". For the intuitive annotations, only the "Y", "N", and "Q" class labels are used, "U" is irrelevant as an intuitive judgment.

The annotations made the challenge a multi-class, multi-label classification task focused on obesity and its co-morbidities. 15 most representative co-morbidities of obesity were considered in this task. The challenge consisted in two tracks, one for the intuitive annotations, and one for the textual annotations. The participating teams had the choice to submit up to three runs for each of the annotations or for both.

The following two tables present the training data sets for each type of annotation, textual and intuitive, for all 16 diseases, obesity with its 15 co-morbidities. The test data consisted in an additional 40% of annotated records.

The evaluation metrics used in the challenge are: precision, recall, and F-measure. Due to high variations in the labels of the data, the macro-averaged F-measure is considered the primary evaluation measure, followed by the micro-averaged F-measure.

[1] http://www.census.gov/

Disease	Class Label				Total
	Y	N	Q	U	
Asthma	93	3	2	630	728
CAD	399	23	7	292	721
CHF	310	11	0	399	720
Depression	104	0	0	624	728
Diabetes	485	15	7	219	726
Gallstones	109	4	1	615	729
GERD	118	1	5	599	723
Gout	90	0	4	634	728
Hypercholesterolemia	304	13	1	408	726
Hypertension	537	12	0	180	729
Hypertriglyceridemia	18	0	0	711	729
OA	115	0	0	613	728
Obesity	298	4	4	424	730
OSA	105	1	8	614	728
PVD	102	0	0	627	729
Venous Insufficiency	21	0	0	707	728

Table 1. Training data sets for textual annotation.

Disease	Class Label				Total
	Y	N	Q	U	
Asthma	86	596	0	NA	682
CAD	391	265	5	NA	661
CHF	308	318	1	NA	627
Depression	142	555	0	NA	697
Diabetes	473	205	5	NA	683
Gallstones	101	609	0	NA	710
GERD	144	447	1	NA	592
Gout	94	616	2	NA	712
Hypercholesterolemia	315	287	1	NA	603
Hypertension	511	127	0	NA	638
Hypertriglyceridemia	37	665	0	NA	638
OA	117	554	1	NA	672
Obesity	285	379	1	NA	665
OSA	99	606	8	NA	713
PVD	110	556	1	NA	667
Venous Insufficiency	54	577	0	NA	631

Table 2. Training data sets for intuitive annotation.

Description of the Methods

We participated in the competition with three systems for both the textual and the intuitive track. In the following paragraphs we will describe the three set-up designs that we used.

Method 1: Bag-of-words

The method represents the well known bag-of-words (BOW) representation technique with frequency feature values. Each record is considered an instance for the classification system.

Features for the BOW representation consist in words delimitated by the following characters: *space, (,) , [,] , . , ' , _*. We do not consider valid features words that have a length smaller than 3. We further reduce the feature space by removing stop-words. For this preprocessing step we used a list of general English stop-words[2] of approximately 700 words.

Method 2: UMLS concepts

In order to work with a representation that provides features that are more general than the words in the abstracts (as used in the BOW representation), we also used a Unified Medical Language System[3] (UMLS) concept representations. UMLS is a knowledge source developed at the U.S. National Library of Medicine (NLM) and it contains a Metathesaurus, a Semantic Network, and the Specialist lexicon for biomedical domain.

The Metathesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts.

UMLS contains over 1 million biomedical concepts, as well as over 5 million concept names which are hierarchically organized. Each unique concept that is present in the thesaurus has associated multiple text string variants (slight morphological variations of the concept). NLM created this knowledge base by

[2] http://www.site.uottawa.ca/~diana/csi5180/StopWords

[3] http://www.nlm.nih.gov/pubs/factsheets/umls.html

unifying hundreds of other medical knowledge bases and vocabularies (around 100 different source vocabularies like MeSH, SNOMED CT, etc.), in order to create an extensive resource that provides synonymy links, as well as parent-child relationships, among single or multi-word concepts. All concepts are assigned at least one Semantic Type from the Semantic Network; this provides a generalization of the existing relations between concepts. There are around 135 Semantic Types in the knowledge base that are linked through 54 relationships.

In addition to the UMLS knowledge base, NLM created a set of tools that allow easier access to the useful information. MetaMap[4] is a tool created by NLM that maps free text to biomedical concepts in the UMLS, or equivalently, it discovers Metathesaurus concepts in text. With this software, text is processed through a series of modules. First it is parsed into components including sentences, paragraphs, phrases, lexical elements, and tokens. For each of the noun phrases that the system finds in the text, variant noun phrases are generated. The variants consist of one or more noun phrases combined together with all their spelling variants, abbreviations, acronyms, synonyms, etc. For each of the variant noun phrases, candidate concepts (concepts that contain the noun phrase variant) from the UMLS Metathesaurus are retrieved and evaluated. The retrieved concepts are compared to the actual phrase using a fit function that measures the text overlap between the actual phrase and the candidate concept (it returns a numerical value). The evaluation is made by also looking at the surrounding context for concepts that are ambiguous. The best of the candidates are then organized according to the decreasing value of the fit function. We used the top concept candidate for each identified phrase in an abstract as a feature.

Figure 1 presents an example of the output of the MetaMap system for the phrase "*to an increased risk*". The information present in the brackets, "Qualitative Concept, Quantitative Concept" for the candidate with the fit function value 861 is the concept used as feature in the UMLS representation.

Another reason to use a UMLS concept representation is the ***concept drift*** phenomenon that can appear in a BOW representation. Especially in the medical domain texts, this is a frequent problem as stated by Cohen *et al*. (2004).

Figure 1. Example of MetaMap system output

Meta Candidates (6)
861 Risk [Qualitative Concept, Quantitative Concept]
694 Increased (Increased (qualifier value)) [Functional Concept]
623 Increase (Increase (qualifier value)) [Functional Concept]
601 Acquired (Acquired (qualifier value)) [Temporal Concept]
601 Obtained (Obtained (attribute)) [Functional Concept]
588 Increasing (Increasing (qualifier value)) [Functional Concept]

New articles that publish new research on a certain topic bring with them new terms that might not match the ones that were seen in the training process in a certain moment of time. Using a more general representation can be a step forward in solving the ***concept drift*** problem.

Method 3: Genia Tagger noun-phrases

Our third representation method uses noun-phrases identified by the Genia tagger[5]. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE[6] abstracts.

Figure 2 presents an example of the output of the Genia tagger for the sentence: "*Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin.*". The tag O stands for Outside, B for Beginning, and I for Inside.

Figure 2. Example of Genia tagger output

Inhibition	Inhibition	NN	B-NP	O	
of	of	IN	B-PP	O	
NF-kappaB	NF-kappaB	NN	B-NP	B-protein	
activation	activation	NN	I-NP	O	
reversed	reverse	VBD	B-VP	O	
the	the	DT	B-NP	O	
anti-apoptotic	anti-apoptotic	JJ	I-NP	O	
effect	effect	NN	I-NP	O	
of	of	IN	B-PP	O	
isochamaejasmin	isochamaejasmin	NN	B-NP	O	
.	.	.	O	O	

The noun-phrases identified by the tagger are considered as features for our representation technique. We applied the following preprocessing steps before defining our set of final features: we removed the following punctuations: *[. , ' ( ) # $ % & + * / = < > [ ] -_]*, we removed stop-words (the same list as for our BOW representation was used),

[4] http://mmtx.nlm.nih.gov/

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[6] http://medline.cos.com/

and we considered valid features only the lemma-based forms of the identified noun-phrases.

As a classification algorithm for our three representation techniques we used the AdaBoost classifier from the Weka (Witten and Frank, 2005) tool[7]. AdaBoost is a classifier that is known to work well with text. It is intended to boost the performance of a nominal classifier, the Decision Stump classifier in our experiments. In the validation experiments performed on the training data sets using 10-fold cross-validation, we also considered the Complement Naïve Bayes (CNB) classifier (Frank and Bouckaert, 2006). Since the AdaBoost results were consistently better for both the textual and intuitive tracks we have decided to take into consideration only the AdaBoost classifier.

Results on the test data

In this section we present the results that we submitted for both tracks in the I2B2 Challenge with the three methods described earlier.

The next tables present the accumulative results for all diseases for the textual and the intuitive track.

Run	**Precision (%)**		**Recall (%)**		**F-Measure (%)**	
	Micro/Macro		**Micro/Macro**		**Micro/Macro**	
BOW	**93.09**	**96.35**	**93.09**	**45.37**	**93.09**	**45.79**
UMLS	90.15	94.45	90.15	43.39	90.15	43.84
NPs	73.55	83.03	73.55	31.96	73.55	32.24

Table 3. Results on the test set for the textual track.

Run	**Precision (%)**		**Recall (%)**		**F-Measure (%)**	
	Micro/Macro		**Micro/Macro**		**Micro/Macro**	
BOW	**92.39**	**94.59**	**92.39**	**60.21**	**92.39**	**60.69**
UMLS	89.01	92.60	89.01	56.74	89.01	57.76
NPs	73.21	78.86	73.21	43.82	73.21	44.33

Table 4. Results on the test set for the intuitive track.

The results marked in bold represent our best scores.

Table 5 presents the statistics for the teams that participated in the competition.

[7] http://www.cs.waikato.ac.nz/ml/weka/

Track	**Precision (%)**		**Recall (%)**		**F-Measure (%)**	
Textual	**Micro/Macro**		**Micro/Macro**		**Micro/Macro**	
Mean	91	75	91	56	91	56
Std Dev	10	17	10	14	10	15
Intuitive						
Mean	91	78	90	60	90	60
Std Dev	9	17	9	6	9	6

Table 5. Statistics for the results of the challenge.

Conclusion

The results that we obtained with our best methods put us in the 18 rank for the intuitive track and 20 for the textual track in the list of 28 participating teams. The best results were obtained by the BOW representation using the AdaBoost classifier.

In our experiments we considered only a frequency feature representation, but we believe that possible improvements could have been obtained when using a binary representation with the CNB classifier. The CNB classifier implements state-of-the-art modifications of the standard Multinomial Naïve Bayes (MNB) classifier for a classification task with highly skewed class distribution.

We also believe that a set-up design that combines the three representation techniques that we used could be a future development and improvement for this task.

References

1. A.M. Cohen, W.R. Hersh and R.T. Bhupatiraju. 2004. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In: Proceedings of the Thirteenth Text Retrieval Conference TREC 2004.
2. Eibe Frank and R.R. Bouckaert. 2006. Naive Bayes for Text Classification with Unbalanced Classes. In the Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, pp. 503-510.
3. Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) Morgan Kaufmann, 2005.