CLASSIFICATION OF FREE TEXT CLINICAL NARRATIVES (SHORT REVIEW)

Olga Kaurova, Mikhail Alexandrov, Xavier Blanco

***Abstract**: The paper is a limited review of publications (1995-2010) related to the problem of classification of clinical records presented in a free text form. The techniques of indexing and methods of classification are considered. We also pay special attention to the description of document sets used in the mentioned research. Finally, we conclude about the perspective research directions related with the topic.*

***Keywords**: Natural language processing, medical corpus, medical diagnostics, automatic medical classification*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

Introduction

For many years natural language processing (NLP) tools have been successfully used to process information in various applications areas. One of such areas is medicine, where the majority of documents are presented in free text form as primary care encounter notes, physical exams, radiology reports, progress notes, clinical histories, etc. The subject under consideration in this article is automatic classification of clinical data. This problem in effect is reduced to medical diagnostics and so, the topic we discuss can be titled as free text based medical diagnostics.

The automatic classification allows to detect incorrect diagnoses, to find particular clinical events in patient records, and to facilitate the exchange of clinical histories between hospitals. In recent years the problem of clinical text classification was under consideration in several publications.

The paper consists of 6 sections. In Section 2 we describe the problems related to application of NLP tools to clinical documents. We also consider statistical measures used in the reviewed articles. Section 3 refers to corpora used in the research. We describe features of the corpora and difficulties of document classification. NLP methods and tools are presented in Sections 4 and 5. Conclusions and short discussion are in Section 6. The paper contains two tables: the 1-st one with corpus data features and the 2-nd one with NLP methods and results of classifications; there is also an appendix with several corpus data samples.

Problem description

Many researchers have attempted to use natural language processing tools, classification and text mining techniques for processing clinical narratives. But their works have centered mostly on clearly defined domains, such as detection of acute bacterial pneumonia from chest X-ray reports [Fiszman, 2000], detection of breast cancer from mammogram reports [Jain, 1997], identification of episodes of asthma exacerbation [Aronow, 1995a], detection of pediatric respiratory and gastrointestinal outbreaks [Ivanov, 2003]. So, the NLP tools used in these researches were based on clearly defined vocabularies related to given domains.

Generally speaking, the main task of NLP in medical applications consists in automatic encoding clinical data from free text form to numerical form. Such a process requires semantic and syntactic information. Output data after such a transformation can be further automatically classified according to specific goals. For example, Aronow et al. in his work identifies episodes of asthma exacerbation. Basically, the problem is to sort a large collection of documents (medical record encounter notes) according to a specified set of characteristics. These characteristics descry the presence of an acute exacerbation of asthma. Then an automatic system performs a three bin sort. One bin contains encounter notes classified as “very probably exacerbation”. The second bin contains “very probably not exacerbation”. And the third bin contains encounters which are uncertain from the point of view of the classifier [Aronow, 1995a].

Automatic assignment of medical codes is one of the NLP results. In healthcare, diagnostic codes are used to group and identify diseases, disorders, symptoms, human response patterns, and medical signs. We can name several coding systems: ICPC (International Classification of Primary Care), ICD (International Statistical Classification of Diseases and Related Health Problems), NANDA (North American Nursing Diagnosis Association) and others. Most of the works considered in the present review deal with ICPC codes. The ICPC is a method that allows for the classification of the patient’s reason for encounter, the problems/diagnosis managed and primary care interventions. These three elements make out the core constituent parts of an encounter in primary care. The ICPC was first published in 1987 by Oxford University Press and is now commonly used in the United Kingdom, the Netherlands, Norway and other countries.

A multi-class classification problem using ICPC is addressed in the work [Røst, 2006]. The authors classify encounter notes using ICPC codes as classification bins. However, they use not the codes themselves (because their number is very large) but their so called chapter values. The ICPC system contains 17 chapters. Thus, encounter notes are classified in 17 classes. Larkey et al. in their work [Larkey, 1996] operated with another coding system – ICD9. It is a more complex code than ICPC because it consists of two parts: a major category and a subcategory. This makes ICD9 more suited for specialized usage in hospitals. Larkey compared three different types of classifiers for automatic code assignment to dictated inpatient discharged summaries. Each possible code served as a category. The problem consisted in calculation of the probabilities that a document belonged to each category from a given list. As a result, a ranked list of codes (categories) for each document was built.

To evaluate the quality of binary classification of medical data the statistical measures “sensitivity” and “specificity” are used. Sensitivity measures the proportion of actual positives which are correctly identified as such, e.g. the percentage of sick people, who are correctly identified as having the condition. Specificity measures the proportion of negatives, which are correctly identified, e.g. the percentage of healthy people, who are correctly identified as not having the condition. In other words, in medical diagnostics sensitivity is the ability to correctly identify those with the disease, whereas specificity is the ability to correctly identify those without the disease. Chapman et al. [Chapman, 2005] used so-called “positive predictive value” besides sensitivity and specificity. It is the proportion of patients with positive test results, who are correctly diagnosed.

To evaluate the quality of categorization, statistical measures “precision” and “recall” are calculated. Just these characteristics are used in information retrieval. Precision is the proportion between relevant retrieved document set and all retrieved documents (the relevant and not relevant ones) and recall is the proportion between the same relevant retrieved document set and all relevant documents (the retrieved and not-retrieved ones).

However, there is a big difference between the typical information retrieval problem and classification of medical data. In information retrieval one aims to provide relevant documents at the top of a belief list. It gives high precision with low level of recall. In classification one aims to classify as many documents as possible, i.e. to achieve high recall, not the high precision [Aronow, 1995b]. The gold standard which is used for evaluation of a given classification procedure is normally prepared by a certified physician or several independent physicians [Chapman, 2005; Fiszman, 2000].

Corpora and their features

Many difficulties in applying NLP methods to medical domain spring from the peculiar character of input data. A vast amount of patient data is available only in free text form: encounter notes, radiology reports, discharge summaries, admission histories, reports of physical examinations, etc. Below we consider some of these types of data in detail.

3.1. Primary care encounter notes

The characteristic features of a corpus made up of primary care encounter notes are: sparseness, brevity, heavy use of abbreviations, many spelling mistakes. This is due to the fact that the notes are normally written during the consultation with a patient when the time is limited. Another feature of such a corpus is that the data varies greatly in style and length. The texts are written by different physicians, each possessing their own manner of registering clinical information.

A dataset of free-text clinical encounter notes and their corresponding manually coded diagnoses is used in [Røst, 2006]. Totally, there are 482,902 unique encounters. The notes in the experimental dataset are coded according to the ICPC-2 coding system. Each encounter consists of a written note of highly variable length and zero or more accompanying codes. 287,868 of the available encounters have one or more ICPC codes. The final goal of the study was the classification of the notes according to the ICPC-2 code. So, in order to avoid ambiguity in the training data all encounters with more than one code were discarded. The final corpus consisted of 175,167 encounter notes. From these document set 2000 documents were selected randomly to be used as a test set, the remaining were used as a training set.

3.2. Medical texts of specific subdomain (e.g. containing specific illness)

A corpus of respiratory encounter notes is presented in the work of Aronow et al. [Aronow, 1995a]. The corpus is divided into two sets: a test and a training collection for automated identification of episodes of asthma exacerbation. The test collection consists of 965 encounter notes of 76 randomly selected asthmatic patients. The training collection consists of 1,368 encounters of other 100 random patients. The corpus is mostly made up of handwritten provider notes, manually entered letter-for-letter by trained inputters. The goal is to sort medical record encounter notes in two groups: those with the evidence of acute exacerbation of asthma and those without it. The corpus was filtered to reduce the number of irrelevant encounter notes. All the notes without the mention of code for asthma or asthma-like conditions (Acute Bronchitis, Bronchiolitis and Bronchospasm) were eliminated. Notes without a definite diagnosis were excluded as well. The resultant corpus numbered 231 texts in the testing collection and 357 in the training collection.

A different approach to creating a corpus was used by Fiszman et al. [Fiszman, 2000]. They dealt with about 15,000 chest x-ray reports produced during a six - month period. All reports were related to acute bacterial pneumonia. From this document set they selected 292 on the following basis: 217 were randomly selected from all the reports of the first three months, while the remaining 75 reports were randomly selected from the following three months from the list of patients with the diagnosis of bacterial pneumonia. Such an artificial way increased the prevalence of pneumonia-related reports in the sample, but at the same time it caused some doubts concerning validity of statistical assessments of classification quality.

3.3. Triage chief complaints, notational texts

Chapman et al. [Chapman, 2005] in their work created a collection of free-text triage chief complaints (TCC) - the earliest clinical data available on most hospital information systems. Triage chief complaints are used to describe the reason for a patient’s visit to an emergency department. In their research the authors used 4700 complaints as a training set and 800 complaints as a test set in order to classify TCCs into 8 syndromic categories. Main characteristics of the corpus data result from the purpose of triage chief complaints: to describe an emergency patient's condition using as short phrases as possible. So, the corpus is made up of short TCC strings that contain a lot of abbreviations, truncations, spelling and punctuation mistakes. This aggravates the problem of automatic medical classification as the data needs to pass through a complicated preprocessing stage in order to be expanded from abbreviated into a more complete form.

A different type of data but with similar "preprocessing" problems was analyzed in the work of Barrows et al. [Barrows, 2000]. The researchers tried to extract relevant diagnosis of glaucoma. A corpus they used was made up of 12,839 ophthalmology visit notes presented in the form of "notational text", a special kind of clinical documentation. Notational texts are typed by physicians during routine patient encounters. Therefore, the corpus data is terse, full of abbreviations and symbolic constructions, some of which may be specific to a medical sub-domain, to an institution, or even a clinician. Statements in notational text are poorly formed according to grammatical construction rules and are considerably lacking in punctuation.

3.4. Discharge clinical notes

Nowadays one of the problems to be solved is encoding medical documents using the ICD-9 code. The reason consists in wide computerization of hospitals. Franz et al. [Franz, 2000] in their work tested three different approaches to automatic disease coding using a German corpus of free-text discharge diagnoses. The corpus is made up of 120000 diagnosis records and the data covers the whole range of clinical medicine. The main problem the researchers faced in their work consisted in low quality texts with spelling errors, ambiguities and abbreviations. Sometimes one phrase included a combination of diagnoses.

In the framework of the automatic assignment of ICD-9 codes Larkey et al. [Larkey, 1996] operated with another corpus of discharge clinical data. Their corpus consisted of 11.599 discharge summaries divided into a training set of 10.902 documents, a test set of 187 documents, and a tuning set of 510 documents. Each summary included several ICD-9 codes (from 1 to 15). The characteristic feature of the corpus is that the summaries vary greatly in length, namely, from 100 to 3000 words per document. Moreover, the notes are written by different doctors, so the data also varies considerably in style.

A challenging problem of detecting possible vaccination reactions in clinical notes was addressed in the work of Hazlehurst et al. [Hazlehurst, 2005a]. The authors created a large corpus, the unusual feature of which consisted in the diverse nature of the data: telephone encounters, emergency department visits and outpatient office visits, including visits for immunization. Telephone encounters are included in the corpus because people often use the nurse advice line to inquire about possible reactions to immunizations converting it into a rich source of immunization-related events. The process of creating and processing the corpus was caused by the necessity to adapt the previously developed NLP-system (MediClass) to the particular goal of the given research. A manually coded training set of 248 patient records was created. Another set of 13657 visits was created, from which 1000 records were used to train the system. The efficiency of the system was finally evaluated on a test set of the remaining 12631 records (26 records were excluded due to corrupted text notes).

3.5. Lexical resource

In the reviewed works both one-word and multiword medical terms are used as keywords. Just these keywords form a parameter space for procedures of classification. However, there is no information on how multiword terms are constructed. In [Yagunova, Pivovarova, 2010] a statistical method for collocation construction is proposed. By a collocation the authors mean any nonrandom co-occurrence of two or more lexical units, specific for either language as a whole or particular genre of texts (corpus). Their method exploits two statistical measures: Mutual Information (MI) and t-score. MI is a coefficient of association strength, while t-score can be understood as a modification of the collocation frequency. The results showed that extracted MI-collocations consisted of such multiword expressions as terminology and nominations; MI-measure proved to be useful for determining subject domain. T-score, in turn, picks out functional grammatical compounds and high-frequency constructions.

One should say that the potential possibilities of classification, any classification, are defined by relations between classes. The closer their characteristics are the lower level of quality we obtain while distributing objects between classes. In case of document classification the closeness between classes is mainly defined by the intersection of lexis related with each class. When classes, e.g. diseases, have absolutely different descriptor lists the quality of classification is the highest: we can avoid any errors. But when lexical resources of each class are similar then one can expect many errors. These extreme cases say about so-called wide domain and narrow domain with respect to classes, which compose this domain.

Unfortunately, the authors of the publications mentioned above did not consider their corpus of medical documents from the point of view of lexis used for disease description. Just for this reason we can say nothing about domains reflected in a corpus: whether they are wide or narrow ones. Taking this circumstance into account could essentially improve the results of classification. We could find only one work where the problem of clustering/classification of documents is considered from the point of view of width and narrowness of a given domain. It is the doctoral dissertation of David Pinto [Pinto, 2008].

In the Table 1 we present the features of the corpora used in the reviewed articles. Appendix contains some examples of clinical records from these corpora.

Table1. Features of the corpora

Study	Domain type	Total number of documents used in primary search	Training set	Test set	Typical features
Røst et al., 2006	Primary care encounter notes	175167	173167	2000	Sparseness, brevity, heavy use of abbreviations, spelling mistakes.
Barrows et al., 2000	Ophthalmology visit notes in the form of “notational text”	12,839			Terseness; abbreviations; specific symbolic constructions; ungrammatical statements; poor punctuation
Aronow et al., 1995	Encounter notes of respiratory care (acute asthma exacerbation)		1368 → 351	965 →231	Handwritten provider notes, manually entered letter-for-letter as written by trained inputters => specific medical abbreviations and terminology.
Fiszman et al., 2000	Chest x-ray reports	15000	292		
Chapman et al., 2005	Triage chief complaints		4700	800	Very short, abbreviations, truncations, spelling and punctuation mistakes
Franz et al., 2000	Free-text discharge diagnoses	120000			Abbreviations; orthographic variations; combination of two diagnosis phrases within one diagnostic statement (typically, a noun phrase with a prepositional phrase)
Larkey et al., 1996	Discharge summaries	11599	10902	187	Vary in length; heterogeneous in linguistic style; much free form text irrelevant to the coding task
Hazlehurst et al., 2005	Post-immunization encounter records		248 + 1000	12631	

Methods

The methods of classification used in the reviewed publications belong to methods of Machine Learning. So, one can find their descriptions in well-known books related with this area [Mitchell, 1997; Bishop, 2006]. The specificity of classification of clinical texts consists in

- preprocessing low quality data (abbreviations, ungrammatical statements, etc)
- necessity to take into account hidden information (relations between descriptors related with diseases)

4.1. Bayesian classifiers

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. Each node in the graph represents a random variable (observable quantities, latent variables, unknown parameters or hypotheses), while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node.

In the Table 2 we present a comparison of several NLP-tools. As one can see, the researchers mostly use Bayesian inference network, achieving good results [Aronow, 1995a, 1995b; Barrows, 2000; Fiszman, 2000; Chapman, 2005; Larkey, 1996]. For example, Aronow [Aronow, 1995a] used the Bayesian inference network in order to make decision with respect to asthma and other adjacent diseases.

4.2. Support Vector Machine

Support vector machine (SVM) is very popular in naturally-scientific applications and nowadays it becomes widely used in the problems of medical document classification. Standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of. This makes the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM was used in the work of Røst [Røst, 2006]. His automatic system was trained on examples using SVM classifier. The achieved accuracy is 49,7%, but this approach is considered as quite promising.

4.3. K-nearest neighbors

k-nearest neighbors (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbors algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. *k* is a positive integer, typically small.

An interesting method was proposed by Larkey et al. [Larkey, 1996]. The authors combined the described k-NN method with Bayesian classifiers and Relevance Feedback. Relevance Feedback has typically been used in information retrieval to improve existing queries. From the retrieved documents the user indicates a set of relevant documents. The original query and terms from the indicated documents are combined to produce a new query, which is better at ranking relevant documents over non-relevant documents. A small set of features is selected separately for each code, and a query is trained for each code. The comparison of different combinations of classifiers (k-NN, BC, RF) shows that the best result is obtained by combining them all.

4.4. Decision trees

Decision tree learning is a commonly used data mining method. It uses a decision tree as a predictive model, which maps observations about an item to conclusions about the item's target value, i.e. predicts the value of a target variable based on several input variables. In tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree invented by Ross Quinlan. It can be summarized as follows: take all unused attributes and count their entropy concerning test samples → choose attribute for which entropy is minimum (or, equivalently, information gain is maximum) → make node containing that attribute.

Aronow et al. in their work considered an ID3 decision tree learning algorithm besides Bayesian network [Aronow, 1995b]. In their research the authors encoded each document as a list of feature-value pairs and passed it to an ID3 decision tree, which returned a probability that the document was a positive instance.

4.5. Lexis based methods

In the work of Zhang et al. [Zhang, 2010] we find a description of an information retrieval method developed to automatically identify qualified patients for breast cancer clinical trials from free-text medical reports – Subtree match. It is an algorithm that finds structural patterns in patient report sentences that are consistent with given trial criteria. The sub-tree model is constructed in three steps. First, each training sentence is parsed by a syntax parser into a parse tree which describes sentence's syntax structure. Second, for each parse tree, all leaves that are keywords are located and from them, a program back-tracks up the tree for three levels to generate a set of subtrees. Finally, all subtrees found in this way are collected and represented as tree regular expressions. Tree regular expressions are used as search models for the given criterion [Zhang, 2010]. This algorithm proved to be very effective, yielding the results of 90%, 49%, 0.63 (Precision, Recall, F-score respectively) for the most complex model in which both manual and automatic keyword generation were used.

A quite novel approach in using keywords is presented in the paper of Catena et al. [Catena, 2008]. The authors elaborated a simple algorithm for automatic classification of clinical encounter notes based on general category descriptions and their comparison. Category description is a list of keyword frequencies reflecting documents of this category and the novelty of the method consists in using so-called positive, neutral and negative keywords. The sign of keywords (+1,0,-1) reflects its contribution to a given category and to its anti-category. Anti-category is presented by all documents, which do not belong to a given category. The problem of categorization is being solved basing on the rule that a keyword is taken into account only if its density in all texts of a given category exceeds its density in all texts of its anti-category. The results evaluated with Purity-measure and *F*-measure are 0.75 and 0.74 respectively.

Table 2 below presents NLP methods and achieved results of the reviewed works.

Table 2: NLP methods and reporting metrics

Study	Tool	Method	Accuracy	Recall	Precision	Specificity	Sensitivity	Positive predictive value
Røst et al., 2006	SVM-Light	Support Vector Machine	49,7%					
Barrows et al., 2000	MEDLEE	Bayesian inference network	95%	90%	100%			
Aronow et al., 1995	(1) INQUERY (2) FIGLEAF	(1) Bayesian inference network (2)ID3 decision tree algorithm			(1)71.7 % (2)80.8 %			

Fiszman et al., 2000	Symtext	Bayesian inference network		95%	78%	85%		
Chapman et al., 2005	Mplus	Bayesian inference network				98%	98%	91%
Franz et al., 2000	MS Access + Visual Basic	(1) Support Vector Machine (2) Heuristic approach	(1) 40% (2) 50,4%					
Larkey et al., 1996	INQUERY	k-NN, relevance feedback, Bayesian independence classifiers		77.6%	57.0%			
Hazlehurst et al., 2005	MediClass							57%
Zhang et al., 2010		Subtree match		49%	90%			

Tools

In this section we present tools used in the reviewed papers. We also inform about other tools, which were not mentioned in the publications but could be useful for automatic medical classification.

SymText (Symbolic Text Processor) is a NLP system, which uses syntactic and semantic knowledge to model the underlying concepts in a textual document [Koehler, 1998]. SymTexts syntactic component contains a parser that uses an augmented transition network grammar and a transformational grammar. SymTexts semantic component includes Bayesian networks that model the relevant medical domain. The system was created at LDS Hospital in Salt Lake City, Utah

MedLEE is a text processor that extracts and structures clinical information from textual reports and translates the information to terms in a controlled vocabulary [MedLEE, http]. Clinical information then can be accessed by further automated procedures. It has been used in radiology, discharge summaries, sign out notes, pathology reports, electrocardiogram reports, and echocardiogram reports, and can readily be ported to other clinical domains. MedLEE was created by Carol Friedman in collaboration with the Department of Biomedical Informatics at Columbia University, the Radiology Department at Columbia University, and the Department of Computer Science at Queens College of CUNY.

INQUERY and **FIGLEAF** were developed in the Center for Intelligent Information Retrieval (CIIR), University of Massachusetts, Amherst. The INQUERY is a text based information retrieval system that uses a probabilistic inference net model [Callan, 1992]. FIGLEAF (Fine Grained Lexical Analysis Facility) is a text classification system based on statistical analysis of semantic features. It uses decision trees derived from examples in a set of training documents [Lehnert, 1995].

Mplus (The Medical Probabilistic Language Understanding System) is a robust medical text analysis tool with a Bayesian network-based semantic model for extracting information from narrative patient records [Christensen, 2002]. The advantage of the system is that its semantic model can be trained in specific domains to adapt to new tasks.

MediClass (Medical Classifier) is a knowledge-based system that automatically classifies the content of a clinical encounter captured in the medical record [Hazlehurst, 2005b]. MediClass accomplishes this by applying a set of application-specific logical rules to the medical concepts that are automatically identified in both the free-text and precoded data elements.

Weka (Waikato Environment for Knowledge Analysis) is a machine learning software written in Java, developed at the University of Waikato, New Zealand [Weka, http]. Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. The system includes several standard data mining tools, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection, which turns it into a useful tool of NLP in medical domain.

RapidMiner (developed at the Artificial Intelligence Unit of the University of Dortmund, Germany) is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics, written in Java [RapidMiner, http]. It provides data mining and machine learning procedures including: data loading and transformation, data preprocessing and visualization, modeling, evaluation, and deployment. It also integrates learning schemes of the Weka and statistical modeling schemes of the R-Project [R, http]

CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of various clusters [CLUTO, http]. It was developed at the University of Minnesota, Minneapolis, USA. CLUTO consists of both a library and stand-alone programs, via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO. Among the key features of the system are: multiple classes of clustering algorithms (partitional, agglomerative, graph-partitioning based), multiple similarity/distance functions (Euclidean distance, cosine, correlation coefficient, extended Jaccard).

Conclusion

The paper contains the review of a number of publications related with processing clinical narratives. By ‘processing’ we mean application of models and methods of computational linguistics for classification of short medical texts presented in free text form.

The review includes:

- analysis of medical corpora from the point of view of their style, volume and domains;
- consideration of methods of classification, which are part of machine learning methods;
- presentation of tools both mentioned in the publications and those could be useful for future applications.

Such a consideration allows to outline the following routes for improving the quality of classification:

- detailed pre-analysis of corpora for revealing linguistic properties of texts, in particularly, whether we deal with wide- or narrow domain;
- application of advanced indexing techniques including word collocations, etc.;
- application of both classification methods and classification technologies including assembling, boosting, etc.

Acknowledgements

As members of ACM's Special Interest Group on Health Informatics (SIGHIT) [http://www.sighit.org] we would like to thank all the group organizers and Dr. Ted Peterson in particular for providing this inexhaustible resource of publications. We also thank Dr. Elena Yagunova for sharing with us her experience and publications in the field of collocation construction.

Bibliography

[Aronow, 1995a] D.B.Aronow, J.R.Cooley, S.Soderland. Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record. In: Proc. of Annual Symposium on Computer Applications in Medical Care. 309-13, USA, 1995.

[Aronow, 1995b] D.B.Aronow, S.Soderland, J.M.Ponte, F.Feng, B.Croft, W.Lehnert. Automated classification of encounter notes in a computer based medical record. In: Proc. of MEDINFO '95 8th World Congress on Medical Informatics, Medinfo, Canada, p. 8-12, 1995.

[Barrows, 2000] R.C.Barrows, M.Busuioc, C.Friedman. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In: Proc. of American Medical Informatics Association Annual Symposium, 51-5, 2000.

[Bishop, 2006] C. Bishop. Pattern Recognition and Machine Learning, Springer, 2006

[Callan, 1992] J.P.Callan, W. B.Croft, S.M.Harding. The INQUERY Retrieval System. In: Proc. of DEXA-92, 3-rd International Conference on Database and Expert Systems Applications, pp. 78-83, 1992.

[Catena, 2008] A.Catena, M.Alexandrov, B.Alexandrov, M.Demenkova. NLP-Tools Try To Make Medical Diagnosis. In: Proc. of the 1-st Intern. Workshop on Social Networking (SoNet-2008), Skalica, Slovakia, 2008.

[Chapman, 2005] W.W.Chapman, L.M.Christensen, M.M.Wagner, P.J.Haug, O.Ivanov, J.N.Dowling, R.T.Olszewski. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine, 33(1), 31-40. 2005.

[Christensen, 2002] L.M.Christensen, P.J.Haug, M.Fiszman. MPLUS: a probabilistic medical language understanding system. In: Proc. of the ACL-02 workshop on Natural language processing in the biomedical domain (BioMed '02), Vol. 3, p. 29-36, Stroudsburg, USA, 2002. (http:// acl.ldc.upenn.edu/W/W02/W02-0305.pdf)

[CLUTO, http] CLUTO: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

[Fiszman, 2000] M.Fiszman, W.Chapman, D.Aronsky, R.S.Evans, P.J.Haug. Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports. American Medical Informatics Association Vol. 7 Num. 6, p. 593-604, 2000.

[Franz, 2000] P.Franz, A.Zaiss, S.Schulz, U.Hahn, R.Klar. Automated coding of diagnoses--three methods compared. Proc. of AMIA Symp. 250-4, 2000.

[Hazlehurst, 2005a] B.Hazlehurst, J.Mullooly, A.Naleway and B.Crane. Detecting Possible Vaccination Reactions in Clinical Notes. In: Proc. of AMIA Annu Symp Proc. 2005; 2005: 306–310.

[Hazlehurst, 2005b] B.Hazlehurst, H.R.Frost, D.F.Sittig, V.J.Stevens. 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, 2005.

[Heinze, 2001] D.T.Heinze, M.L.Morsch, and J.Holbrook. Mining Free-Text Medical Records. In: Proc. AMIA Symp. 2001; 254–258, 2001.

[Hofmans-Okkes, Lamberts, 1996] I.M.Hofmans-Okkes, H.Lamberts. The International Classification of Primary Care (ICPC): new applications in research and computer-based patient records in family practice. Family Practice; Vol. 13, No 3, p. 294-302, 1996.

[Hripcsak, 2002] G.Hripcsak, J.Austin, P.Alderson & C.Friedman. Use of natural language processing to translate clinical information from database of 889,921 chest radiographic reports. Radiology (224), 157-163, 2002.

[Ivanov, 2003] O.Ivanov, P.Gesteland, W.Hogan, M.B.Mundorff, M.Wagner. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. In: Proc. of AMIA Annual Fall Symposium; p. 318—22, 2003.

[Jain, Friedman, 1997] N.L.Jain, C.Friedman. Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports. Proc. of AMIA Annu Fall Symp., p. 829-33, 1997.

[Koehler, 1998] S.B. Koehler. Symtext: A Natural Language Understanding System For Encoding Free Text Medical Data. Doctoral Dissertation, Department of Medical Informatics, University of Utah, 1998

[Larkey, Croft, 1996] L.S.Larkey and W.B.Croft. Combining classifiers in text categorization. In: Proc. of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96), pp. 289-97, Zurich, Switzerland. ACM Press, 1996.

[Lehnert, 1995] W.Lehnert, S.Soderland, D.Aronow, F.Feng, A.Shmueli. Inductive text classification for medical applications. In: Journal of Experimental and Theoretical Artificial Intelligence, 7:49-80, 1995.

[MedLEE, http] MedLEE: http:// lucid.cpmc.columbia.edu/medlee/

[Mitchell, 1997] T.Mitchell. Machine Learning, McGrow Hill, 1997.

[Pinto, 2008] D. Pinto, On Clustering and Evaluation of Narrow Domain Short-Text Corpora. Doctoral Dissertation, Polytechnic University of Valencia, Spain, 2008

[R, http] R-project: http:// www.r-project.org

[RapidMiner, http] RapidMiner: http://rapid-i.com

[Røst, 2006] T.B.Røst, O.Nytro, A.Grimsmo. Classifying Encouter Notes in the Primary Care Patient Record. In: Proc. of the 3-rd Intern. Workshop on Text-Based Information Retrieval (TIR-06). Univ. Press, p. 5-9, 2006.

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

[Yagunova, Pivovarova, 2010] E. Yagunova, L. Pivovarova. The Nature of Collocations in the Russian Language. The Experience of Automatic Extraction and Classification of the Material of News Texts // Automatic Documentation and Mathematical Linguistics, 2010, Vol. 44, No. 3, pp. 164–175. © Allerton Press, Inc., 2010. Original Russian Text © E.V.Yagunova, L.M.Pivovarova, 2010, published in Nauchno Tekhnicheskaya Informatsiya, Seriya 2, 2010, No. 6, pp. 30–40.

[Zhang, 2010] J.Zhang, Y.Gu, W.Liu, T.Zhao, X.Mu, W.Hu. Automatic Patient Search for Breast Cancer Clinical Trials Using Free-Text Medical Reports'. In: Proc. of the 1-st ACM International Health Informatics Symposium. New York, USA., 2010.

[Zhou, 2006] X.Zhou, H.Han, I.Chankai, A.A.Prestrud, A.Brooks. Approaches to Text Mining for Clinical Medical Records. In: Proc. of the 21-st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Computer Applications in Health Care (CAHC 2006), Dijon, France. 2006.

Appendix

Some samples of clinical corpus data

Sample 1: A typical encounter note from [Røst et al., 2006]:

"Inflamed wounds over the entire body. Was treated w/ apocillin and fucidin cream 1 mth. ago. Still using fucidin. Taking sample for bact. Beginning tmnt. with bactroban. Call in 1 week for test results".

Sample 2: An encounter note for acute asthma exacerbation [Aronow et al., 1995]:

"G100 ASTHMA

COUGH & WHEEZE X1-2D HX PNEU 4/92 AFEB

ACTIVE RR=48 W/MOD RETRAX CHEST

DIFFUSE EXP WHEEZING & RHONCHI ONLY

SL. CLEARING AFTER 2 NEBS 02 SATS 93->

HOSP ER."

Sample 3: Notational text from [Barrows et al., 2000]:

"3/1198 IPN

SOB & DOE$

VSS, AF

CXR 3LLL ASD no A

WBC IIK

SIB Cx 69GPC c/W PC, no GNR

DIC Cef – PCNIV”

Authors' Information

***Olga Kaurova** – Saint Petersburg State University (Department of Theoretical and Applied Linguistics - graduated in 2009); Autonomous University of Barcelona (International Master in “Natural Language Processing & Human Language Technology” - graduated in 2010; PhD program “Lenguas y Culturas Románicas” - current), 08193 Bellaterra (Barcelona), Spain;*

e-mail: kaurovskiy@gmail.com

Major Fields of Scientific Research: automatic medical classification, sentiment analysis

***Mikhail Alexandrov** – Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: MAlexandrov@mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

***Xavier Blanco –** Cathedratic University Professor (Full Professor), fLexSem Research Laboratory (Fonètica, Lexicologia i Semàntica), Department of French and Romance Philology, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: Xavier.Blanco@uab.cat*

Major Fields of Scientific Research: lexicology, lexicography, machine translation