Detecting Automatic Patterns of Stroke Through Text Mining

Miguel Vieira, Filipe Portela(✉), and Manuel Filipe Santos

Algoritmi Research Center, University of Minho, Guimarães, Portugal
{cfp,mfs}@dsi.uminho.pt

Abstract. Despite the volume increase of electronic data collection in the health area, there is still much medical information that is recorded without any systematic pattern. For instance, besides the structured admission notes format, there are free text fields for clinicians' patient evaluation observation. Intelligent Decisions Support Systems can benefit from cross-referencing and interpretation of these documents. In the Intensive Care Units, several patients are admitted daily, and several discharge notes are written. To support real-time decision-making and to increase the quality of its process, is crucial to have all relevant patient clinical data available. Since there is no writing pattern followed by all medical doctors, its analysis becomes quite difficult to do. This project aims to make qualitatively and quantitatively analysis of clinical information focusing on the stroke or cerebrovascular accident diagnosis using text analysis tools, namely Natural Language Processing and Text Mining. Our results revealed a set of related words in the clinician' patient diaries that can reveal patterns.

Keywords: Medical information · Admission notes · Intelligent Decisions Support Systems · Intensive Care Units

1 Introduction

The medical history of patients is typically documented in clinical notes that are stored in the Electronic Medical Records of the respective healthcare organisation, and contain information about admission, diagnosis, laboratory examination results, potential operations, medication, among others. Clinicians write these notes without following a specific writing pattern, constituting unstructured text that difficult information retrieval. Currently, there are computational tools that enable the analysis of free-text, in different languages. This analysis enables the automatic extraction of valuable information about the patient, which might be ignored. Interestingly, it is possible to find patterns and establish standards for specific clinical events of interest.

The purpose of this article is to analyse unstructured text, finding patterns and thereby facilitating the understanding of the information in such clinical events, in this particular case strokes. Therefore, we have used Natural Language and Text Mining (NLP) algorithms, in order to find patterns and word networks in the admission notes relatively to cerebrovascular accidents. With this type of analysis, the physicians can take decisions more accurately and effectively.

P. Cortez et al. (Eds.): INTETAIN 2018, LNICST 273, pp. 58–67, 2019.
https://doi.org/10.1007/978-3-030-16447-8_6

The Centro Hospitalar Hospital de Santo António provided the dataset, which contains information regarding admission to Intensive Care. This work is divided into four main parts: (i) Background part, where the state of the art is presented; (2) Material and Tools section, which describes the various tools used to elaborate this work, namely an explanation of the KH Coder tool as well as the models that are used; (3) Data Study section for description of data treatment used in this work; (4) Results, Discussion and Future Work sections in which the results of this work are described, as well as its discussion and future work perspectives.

2 Background

2.1 Natural Language Processing

The NLP is a computerised approach based on a set of theories and technologies that allows analysing texts, meaning that computers are used to understand and manipulate the language of a specific text or idiom. An ideal NLP system should be capable of Paraphrase inserted text, translate the text in another type of language, questioning the content of the text, as well as to be capable of deducting about the text [1].

In the medical informatics field, there is a long-time concern with medical language. The data about patients processes are non-numeric and formulated almost exclusively within the constructions of natural language [2]. These constructs were identified as syntactic and semantic constructs origins, becoming important in the development of the Systematic Nomenclature of Multifaceted Pathology (SNOP), later known as SNOMED, and currently SNOMED International (SNOMED III). The possibility of automatic coding pathologies and diagnostic reports in SNOP was a success. Nevertheless, researchers from around the world, such as Canada and the United States, continued to work on the automated indexing of Natural Language clinical reports in SNOMED codes [3]. In 2002 a partnership with SNOMED and CTV3 gave rise to a new version of SNOMED called SNOMED CT. The barrier between the use of different terminologies or international coding systems are smaller, and that data can be presented in various ways, depending the purpose, for example the clinical records presented through SNOMED CT can be processed and presented in different ways, to support direct patient care, clinical audit, research, epidemiology, management and service planning [4].

2.2 Text Mining

Text Mining, also known as Text Data Mining, is a process of extracting useful information patterns from unstructured texts or documents, being an extension of Data Mining. However, is more complicated that Data Mining, since it handles unstructured text data. It brings together a set of various disciplines such as text analysis, information extraction, categorisation, visualisation, database technologies, Machine Learning and Data Mining [5].

Typically, the Text Mining tasks include the following research activities [6]:

- Text categorisation: associate texts to categories;
- Text clustering: groups the texts by categories;
- Sentiment analysis: understand the tone of the text;
- Entity Relational Modelling: summarise the texts and discover relationships between the entities described in the text;

Text Mining, inspired by Data Mining, refers to the process of Knowledge Discovery in Text, known by the acronym KDT. It consists of obtaining information from a natural language text [7]. Knowledge discovery is defined as an implicit and non-trivial extraction of previously unknown data that may be useful. There are two parts regarding Knowledge Discovery; one part consists in the application of statistical analysis techniques and Machine Learning to find patterns on knowledge bases, where the other part focuses on providing them with a guided use for data exploitation [8].

2.3 Admission Notes

An admission note is part of a medical record documenting a patient's condition, including medical history, physical exams, and justification of patient admission to a particular hospital facility, as well as initial instructions to begin patient treatment. Besides these functions, admission note can also have additional notes of the service, progress notes SOAP (Subjective, Objective, Assessment and Plan, this is a method of documentation employed by health care providers), pre-operative, operational, post-operative, procedure and delivery notes, postpartum grades and discharge grades. The admission criteria may vary depending on the area in which the user is admitted. For example, admission in Pediatric Intensive Care is primarily planned for patients without therapeutic limitations, with functional instability of one or more organs requiring monitoring or treatment that cannot be performed outside the CIPE (Pediatric Intensive Care Services) [9].

2.4 Strokes

Approximately 15,5 million cardiovascular deaths occur every day [10]. Strokes are among the leading causes of death and disability in the developed world. In the United States, approximately 500,000 people have a new or recurrent stroke each year. Of these, 150,000 die yearly of stroke [11]. The brain is fully responsible for intelligence, personality, mood and characteristics that individuate us and lead our fellow humans to recognise us as humans. The loss of brain function can be dehumanising, making us dependent on others. Moreover, what could be worse than the sudden inability to speak, to move a limb, to stand, to walk, to see, to read, or to become seriously incapable of understanding spoken language, writing, thinking clearly or not even have the ability to remember things? The Loss of functions is often instantaneous and entirely unpredictable; the damages can be transitory or permanent, mild or devastating [12].

The World Health Organization (WHO) defines a cerebrovascular accident as a focal (or global) neurological impairment that suddenly occurs with symptoms persisting beyond 24 h, or leading to death, with probable vascular origin. Many of the

patients who survive have physical, sensory and cognitive sequelae. In a synthesised way, a cerebrovascular accident happens when there is a sudden interruption of cerebral blood flow [12]. Approximately 85% of cerebrovascular accidents are ischemic and 15% haemorrhagic being 10% of intraparenchymal haemorrhage (IPH) and 5% of subarachnoid haemorrhages (SAH).

2.5 Related Work

INTCare is a Clinical Decision Support System (CDSS), based on knowledge discovery in databases (KDD), and on agent-based paradigms with the goal of helping medical decision-making. INTCare is a system that helps clinicians make decisions by detecting patient conditions through continuous updates on their health status and applying the predictive model to predict possible failures that may occur in the next day. INTCare also performs up-to-date maintenance on the probability of death used in an end-of-life decision process. Also, the INTCare also assesses the evolution scenarios of the patient's condition, allowing medical doctors to compare the consequences of different medical procedures [13].

In recent years, several types of evaluations have been developed with the objective of estimating hospital mortality in an ICU. In this study, they predicted one-month mortality related to chronic kidney disease using the Medical Information Mart for Intensive Care III (MIMIC III) database. Also, they observed the improvement in predictive performance and the interpretability of the basic model used in the ICU, for a more complex model using simple resources such as unigrams or bigrams, advanced features, as well as extractions of nursing notes. The primary focus was nursing notes, in which patients who died within the first 24 h of admission and notes that were not updated were excluded. In this study, they observed improvements in the predictive performance and interpretability of predictive models based on new resources extracted from the notes collected in nursing EMRs. More precisely, they predicted one-month mortality at the end of 24 h spent in the ICU in patients with chronic kidney disease (CKD) [14].

3 Material and Tools

3.1 Material Used

The tools used for this work are Microsoft SQL Server 2014, Microsoft Excel, and KH Coder.

3.2 KH Coder

KH Coder is freeware highly utilised for content mining, supporting Japanese, English, French, German, Italian, Portuguese and Spanish etymological information. It works only with .txt files with structured or unstructured text and enables to observe which terms are most used, grouping the terms in the cluster, see the terms frequency and the associations of the word. Individually, it can contribute factual examination co-event system hub structure, computerised arranging guide, multidimensional scaling and comparative calculations [15].

In this work, it will be used some commands through KH coder tool. This analysis was carried out through three models, which are Themes Frequency, Self-Organizational Map and Word Associations. The Themes Frequency model uses an algorithm that can find out which words appear most often in the document and this is done through a function. The function *tf(d, t)* is obtained by dividing the frequency of the word *t* in document *d* by its length, where the length of document *d* is the number of words contained in the document, i.e., the number of morphemes.

$$tf(d,t) = Frequency\ of\ the\ word\ t\ in\ the\ document\ d / Length\ of\ the\ document\ d. \tag{1}$$

The Self Organization map enables to explore associations between words through Euclidean distances; the word frequency (adjusted frequency) is standardised before the distance calculation. This ensures the distance of the words is calculated based on occurrence patterns, rather than on whether each word appears frequently or not.

The Word Associations analysis displays a chart where the words closely associated with each other are connected with lines, while words that define the search condition are enclosed in double rectangles. This command enables a way to find words that are closely associated with a specific word, as well as words that are closely associated with a specific code. The possible combinations for the target word are calculated using the Jaccard coefficients [16].

3.3 Dictionary

In order to work with topics of words, a dictionary had to be created for grouping terms into themes which allows a more focused analysis of the document. Anyone can make their own dictionary to analyse the group themes that are related to the project. This dictionary is a group of Portuguese terms that are related to the data, and it was created manually after a deep analysis of the dataset. The purpose of the dictionary is to group words that have the same meaning into topics, making the analysis more accurate and relevant. Without the dictionary, we would have an individual analysis of each word and of words with the same meaning. By grouping the words, the analysis becomes more complete and easier to understand. For example:

*Alteration
alterations | changes | variation | adjustment |
*Negative
no | without | negative |

4 Data Study

The individual patient data used in this analysis were anonymised and provided as part of a partnership with the Hospital Center of Porto, Santo António Hospital.

The dataset has three columns:

- ID: that contains the ID of the patient;
- ENQUADRAMENTO10: the medical history record of the patient, as well as the reasons why the patients are being admitted to the ICU;
- DIAGNÓSTICO: that contains the final patient's diagnosis;

This dataset contains a significant amount of patient's information, so it was needed to treat data, for example deleting all the nulls and extracting only the patient's information with stroke diagnosis. The original dataset has 3363 records, and the dates of these records are between 2010 and April of 2018. The dataset has two types of data format, int and string.

5 Results

Data analysis performed in this work leads to several results. However, it should be noted that these analyses must be interpreted by specialised clinicians, that have the necessary knowledge to make this analysis useful. Using Themes Frequency, that identify the words that appear more frequently in the document, we could observe that the term "artery" appears 45 times, "cerebral" appears 29 times and "right" appear 42 times. With this analysis, we can already have a notion of which are the most important and relevant topics for the study. There is an example of the frequency is presented in Table 1:

Table 1. Frequency topics

Codes	Frequency	Percent
*artery	45	11.45%
*left	34	8.65%
*cerebral	29	7.38%
*right	42	10.69%
*segment	19	4.83%
*thrombus	24	6.11%
*carotid	22	5.60%
*thrombectomy	18	4.58%
*infarct	14	3.56%
*internal	13	3.31%
*cat	19	4.83%
*hemiparesis	13	3.31%
*occlusion	14	3.56%
*hypertension	15	3.82%
*medium	20	5.09%
*neurological	10	2.54%
*admit	0	0.00%
*nihss	12	3.05%
*territory	11	2.80%

(*continued*)

Table 1. (*continued*)

Codes	Frequency	Percent
*previous	8	2.04%
*stroke	11	2.80%
*basilar	10	2.54%
*catheter	8	2.04%
*distal	10	2.54%
*degree	9	2.29%
*line	0	0.00%
*sedate	0	0.00%
*present	0	0.00%
*arterial	9	2.29%
*eye	8	2.04%

A Self-Organizational Map with six defined clusters presents as follows in Fig. 1:

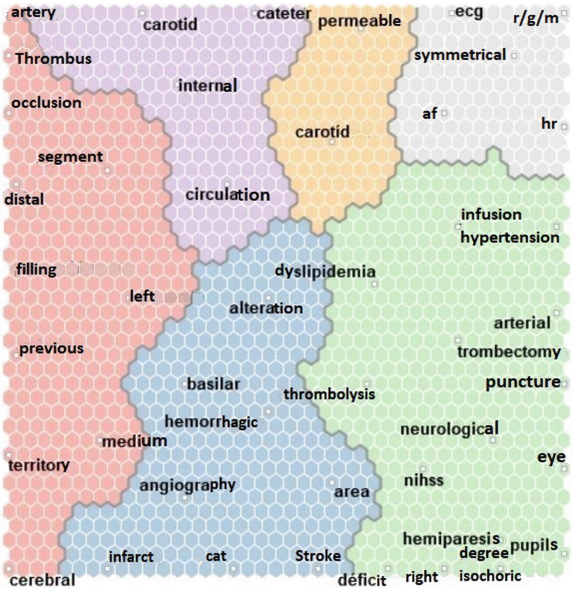

Fig. 1. Self-organization map

Next, we aim to use this command that enables to explore associations between words by creating a self-organising map.

In Fig. 1, it is observed six clusters, and they have several topics inside. The topics are related to the topics inside of the same cluster, and the different colours depicted in the image represent the clusters. This analysis allows to show (through the admission notes where it contains the medical history of all the patients who were hospitalised due

to strokes) what were the symptoms responsible for the hospitalisation of the patients in ICU, and thus group them in clusters.

Interestingly, it is possible to verify that there are six different types of patients diagnosed and admitted to the ICU with a cerebrovascular accident. The Self Organization Map can be modified, but in this way, shows a general view of the document:

- In the orange cluster the topics inside are thrombus, occlusion, segment, distal, filling, left, previous, medium, territory and cerebral.
- In the purple cluster, the topics inside are an artery, carotid, internal, circulation and catheter.
- In the blue cluster, the topics inside are an alteration, basilar, haemorrhagic, angiography, area, infarct, computed axial tomography (cat) and stroke.
- In the yellow cluster, the topics inside are permeable and carotid.
- In the green cluster, the topics inside are an infusion, hypertension, dyslipidaemia, arterial, thrombectomy, puncture, neurological, eye, nihss, hemiparesis, degree, deficit, isochoric, right and pupils.
- In the grey cluster, the topics inside are electrocardiogram (ecg), vesicular murmur (r/g/m), symmetrical, arterial fibrillation (af) and heart rate (hr).

If the patient presents a set of symptoms of one of the clusters it is most likely that he or she had a stroke, or it will have in the future, therefore if a patient does not belong to any of these clusters, most likely he has not suffered, or will not suffer from any type of cerebrovascular accident.

A Word Association with the term "artery" presents as follows (Fig. 2):

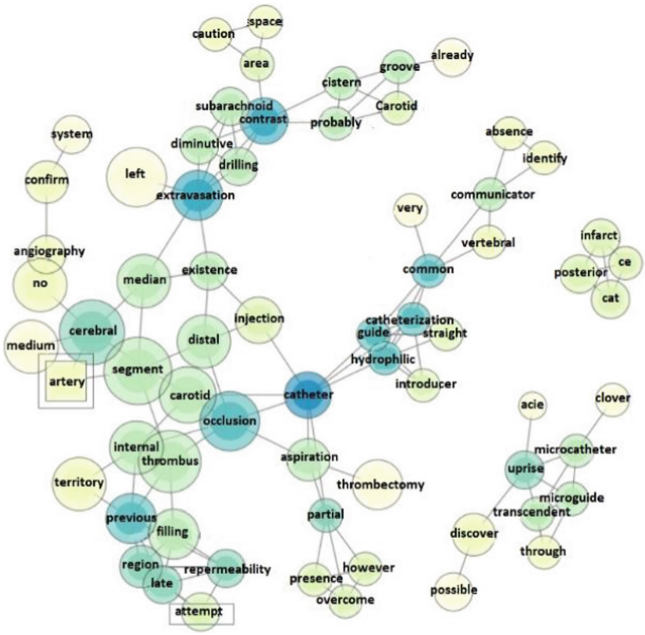

Fig. 2. Word Association with the term artery

Here it can be seen the Word Association, where the chosen word was, “artery” and through that word, it can be seen the network presented in Fig. 2. The chosen word was “artery”, because it was the word that appeared most often in the patient’s admission notes (as shown in Table 1), thus indicating that it is one of the essential words regarding stroke, since strokes occurs when there is a clogging or rupture of the arteries (arteries of the brain).

That shows the words that are associated with artery term. With this technique, it is possible to discover what are the different types of words that occurred with any term that is chosen. For example, the word artery, in the text, it appears associated many times with cerebral as with carotid because cerebral and carotid are two types of the artery. However, the word cerebral appears connected with the word middle that’s because it is also a type of artery, called middle cerebral artery. With this type of analysis is shown associations of any words that appear in the admission notes and obtain relations with other words that could go unnoticed, to discovering new patterns related to strokes.

6 Discussion and Future Work

As shown in the results, through a simple word, or a set of words and their relations several conclusions can be obtained. To this conclusion be more precise and better understood, its necessary to be performed for professionalised people like clinicians with the expert knowledge.

In the first analysis, Table 1, it can be seen what the themes that appear more often are and how they appear in the dataset.

In the second analysis, Fig. 1, it is already seen a qualitative dataset analysis. This analysis enables to divide the terms that occur more frequently by the cluster. As it can be seen in Fig. 1, for example, one of the clusters has the terms “mv” that means vesicular murmurs, “electrocardiogram”, “symmetrical”, “atrial fibrillation”, “hr” that means “heart rate”. These are the terms grouped in a type of admission notes for an ICU. Therefore, by the contrary logic, patients who do not belong to any of these clusters, probably will not suffer from any cerebrovascular accident.

In the Word Association analysis, Fig. 2, shows the relationships of words on a specific chosen topic. This type of analysis can be beneficial to see what are the events, or even diseases, that are related to a specific word, in this case, the chosen word was “artery”.

With this work, it is possible to conclude that through unstructured text, it can be created and discovered patterns, which can be an advantage to make the clinical decision more accurate and faster. Indeed, this type of work can be adapted to other areas of medicine. This work consisted primarily of choosing a particular information type from a global dataset and the creation of a dictionary to allow the various analysis made.

The future work in this field could be focused on exploring a more general and complex dataset, with more information and with more types of diagnoses. After an in-depth analysis with more information, it will be possible to create models of text

mining prediction to automatically predict what the patient may have in the future, for example, diseases.

Acknowledgements. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. This work is also supported by the Deus ex Machina (DEM): Symbiotic technology for societal efficiency gains - NORTE-01-0145-FEDER-000026.

References

1. Chowdhury, G.: Natural language processing. Annual Review of this is an author-produced version of a paper published in The Annual Review of Information Science and Technology ISSN 0066–4200. This version has been peer-reviewed, but does not. Annu. Rev. Inf. Sci. Technol. **37**, 51–89 (2003)
2. Aw, P.: Medicine, computers, and linguistics. Adv. Biomed. Eng. **3**, 97–140 (1973)
3. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. J. Am. Med. Inform. Assoc. **1**(2), 142–160 (1994)
4. SNOMED: SNOMED International, no. Dec 2010 (2006)
5. Tan, A.-H.: Text mining: the state of the art and the challenges. In: Proceedings of PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, vol. 8, pp. 65–70 (1999)
6. Truyens, M., Van Eecke, P.: Legal aspects of text mining. Comput. Law Secur. Rev. **30**(2), 153–170 (2014)
7. Zhao, Y.: Text mining. In: R Data Mining, pp. 105–122 (2013)
8. Feldman, R., Dagan, I.: Knowledge discovery in textual databases (KDT). In: International Conference on Knowledge Discovery and Data Mining, pp. 112–117 (1995)
9. Pedi, C.I.: Critérios de admissão no Serviço de Cuidados Intensivos Pediátricos. pp. 2–3 (2014)
10. Part I: General Considerations, the Epidemiologic Transition: Clinical Cardiology : New Frontiers Global Burden of Cardiovascular Diseases, no. C (2001)
11. De Magalhães, R., De Oliveira, C., Augusto, L., De Andrade, F.: Artigos Acidente vascular cerebral, vol. 8, no. 3, pp. 280–290 (2001)
12. Alves, C.: Determinantes da capacidade funcional do doente após acidente vascular cerebral (2011)
13. Gago, P., Santos, M.F., Silva, A., Cortez, P., Neves, J., Gomes, L.: INTCare: a knowledge discovery based intelligent decision support system for intensive care medicine. J. Decis. Syst. **14**(3), 241–259 (2005)
14. Kocbek, P., Fijačko, N., Zorman, M., Kocbek, S., Štiglic, G.: Improving mortality prediction for intensive care unit patients using text mining techniques, pp. 2–5 (2012)
15. Gowri, S., Anandha Mala, G.S.: Efficacious IR system for investigation in digital textual data. Indian J. Sci. Technol. **8**(12), 43102 (2015)
16. Higuchi, K.: KH coder. Ref. Man. 99 (2016). http://khcoder.net/en/manual_en_v2.pdf