Towards automated classification of intensive care nursing narratives

Marketta Hiissa a,b,*, Tapio Pahikkala a,c, Hanna Suominen a,c, Tuija Lehtikunnas d,e, Barbro Back a,b, Helena Karsten a,c, Sanna Salanterä d, Tapio Salakoski a,c

a Turku Centre for Computer Science, Joukahaisenkatu 3-5 B, 20520 Turku, Finland
b Åbo Akademi University, Department of Information Technologies, Turku, Finland
c University of Turku, Department of Information Technology, Turku, Finland
d University of Turku, Department of Nursing Science, Turku, Finland
e Turku University Hospital, Turku, Finland

Article info

Article history:
Received 29 December 2006
Received in revised form 20 March 2007
Accepted 28 March 2007

Keywords:
Computerized patient records
Intensive care
Natural language processing
Nursing
Nursing records

Abstract

Background: Nursing narratives are an important part of patient documentation, but the possibilities to utilize them in the direct care process are limited due to the lack of proper tools. One solution to facilitate the utilization of narrative data could be to classify them according to their content.

Objectives: Our objective is to address two issues related to designing an automated classifier: domain experts’ agreement on the content of classes Breathing, Blood Circulation and Pain, as well as the ability of a machine-learning-based classifier to learn the classification patterns of the nurses.

Methods: The data we used were a set of Finnish intensive care nursing narratives, and we used the regularized least-squares (RLS) algorithm for the automatic classification. The agreement of the nurses was assessed by using Cohen’s κ, and the performance of the algorithm was measured using area under ROC curve (AUC).

Results: On average, the values of κ were around 0.8. The agreement was highest in the class Blood Circulation, and lowest in the class Breathing. The RLS algorithm was able to learn the classification patterns of the three nurses on an acceptable level; the values of AUC were generally around 0.85.

Conclusions: Our results indicate that the free text in nursing documentation can be automatically classified and this can offer a way to develop electronic patient records.

© 2007 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

During the past years, health-care providers have been changing paper-based patient records to electronic ones. This has, on the one hand, made more data available on each patient, but on the other hand, also offered new possibilities to utilize the gathered data. However, the effects of this switch have not only been positive. It has been found that electronic charting may not provide nurses with more time for tasks unrelated to manipulating data [1,2] and that electronic systems support nurses in gathering information, but not in the active utilization of it [3].

In Finland, most of the intensive care nursing documentation, i.e. documentation about the state of the patient, goals for
care, nursing interventions and evaluation of delivered care, is recorded as free text notes, i.e. narratives, and there is no common practice about how the narratives are written. In some units, paragraph names such as breathing, hemodynamics, elimination and consciousness are used to guide the documentation, but in general, each nurse uses her/his own style. Due to this, incoherence between one document and another is almost unavoidable, which in turn makes it more difficult for example to find relevant information from the records, and thus complicates the utilization of gathered narratives. The active utilization of narrative data is problematic in particular when the patient stays in the ward for several days, and the amount of documentation is large.

Narratives, in addition to serving nurses’ information needs about the state of the patient and care given, also serve as basis for, e.g. writing the discharge report when the patient is discharged from the Intensive Care Unit (ICU) to another ward, and when defining the caring needs and nursing intensity of the patient. The classification used in Finland to define the patient’s caring needs is the Oulu Patient Classification, OPC [4,5], that makes a distinction between six different areas of intensive care nursing. Nurses give scores to each patient in compliance with the categories based on the severity and the degree of difficulty of the problems the patient is suffering from. Based on the given scores, patients are divided into five nursing intensity categories. If it would be possible to systematically classify the written narratives so that they correspond to the six areas used in the OPC, both the defining of the nursing intensity and the writing of the discharge report would become easier, and it would also be easier to search for some specific information needed.

Our approach to facilitate the utilization of narratives is to develop automatic tools that classify texts, and thus make it easier to utilize the information the narratives include. In the medical domain, classification has recently been applied, e.g. to classifying texts such as chief complaint notes, diagnostic statements, and injury narratives into different kinds of syndromic, illness and cause-of-injury categories [6–9]. However, studies applying natural language processing to clinical contents are still relatively rare [10], and there is little research on the automated processing of nursing narratives [11].

In this paper, we use a machine-learning approach, i.e. an algorithm that learns the classification patterns directly from pre-classified data, to classify Finnish intensive care nursing narratives. We address two issues related to designing a classifier: domain experts’ agreement on the content of the classes into which the data are to be classified, and the ability of the classification algorithm to perform on an acceptable level. The classes used in this study are Breathing, Blood Circulation and Pain, and the machine-learning algorithm is the Regularized Least-Squares Algorithm.

2. Material and methods

2.1. Material

The data we used were a set of Finnish intensive care nursing narratives. The documents were gathered in the spring of 2001 from 16 intensive care units, two or three documents per unit. In total, we had 43 copies of patient records including nursing notes written down during one day and night. Nurses wrote the notes during the care process; in intensive care, the documentation always takes place at the patient’s bedside. The language of the notes was Finnish.

The research permission for the study was applied and achieved according to the Finnish laws and ethical guidelines for research. We applied for: (a) a permission from each participating hospital authority that has the right to acknowledge the permission, and (b) a statement about the ethics of the study from an ethical committee for health research. The gathered documents included only nursing notes written down during one day and night, and the documents were anonymized, i.e. no patient identification information was given to the researchers. Within our research group, only one nurse handled the original documents, respecting the professional secrecy. The other researchers had access to a file in which the narrative statements from the individual documents were aggregated, but the recognition of individual patients was impossible. Thus, the anonymity of the patients was preserved during the whole research process.

In the gathered data, the style of the documentation varied from one nurse to another: some had written short sentences such as “Hemodynamics ok. Very tired.” whereas others had written long sentences in which different matters were separated with commas, e.g. “Patient has extra systoles, EKG control tomorrow morning, became bluish when turning him around and the oxygen saturation decreased, with extra oxygenation the situation improved, we made his sedation deeper.” In order to standardize the style of the documentation, we divided the long sentences into smaller pieces consisting of one matter or thought. This was done manually by one of the authors with nursing experience, and resulted in 1363 pieces, with the average length of 3.7 words.

2.2. Methods

In this section, we firstly discuss the classes into which the data was classified, and then we describe how the nurses’ agreement on the content of the classes was measured. We also introduce the Regularized Least-Squares Algorithm, and describe how the performance of the algorithm was measured.

2.2.1. Classes

We classified the data according to classes Breathing, Blood Circulation and Pain, which represent a subset of the content of intensive care nursing. The OPC classification used to assess patients’ caring needs and nursing intensity, divides the intensive care nursing into six areas. These areas are planning and co-ordination of care; breathing, circulation and symptoms of diseases; nutrition and medication; personal hygiene and excretion; activity, movement, sleep and rest; as well as teaching, guidance in care, follow-up and emotional support. The three chosen issues, Breathing, Blood Circulation and Pain, are all assessed within the Breathing, Circulation and Symptoms of Diseases category [4].

1 The citations were translated from the notes written in Finnish. During the study, the original notes in Finnish were used.
Although assessed within the same category in the OPC, the chosen classes represent different aspects of nursing work. Breathing and blood circulation are vital functions, the maintenance and monitoring of which is the first priority of intensive care nursing, and there are different kinds of devices to assist in the monitoring of them [12]. Pain, on the other hand, has to be monitored mainly on the basis of different kinds of behavioural and physiological indicators, which is due to patients being incapable of communicating because of their critical illnesses [13]. We wanted to investigate if the different nature of the documentation of pain had any effect on the performance of the automated classifier.

2.2.2. Nurses’ agreement on the content of the classes

In order to assess domain experts’ agreement on the content of the three classes, we asked three nurses to manually classify the data, independently of each other. The agreement was measured in order to get an insight on how similar opinions nurses have on what kind of statements should be included in the chosen classes. The nurses were advised to label each text piece they considered to contain useful information given the specific class. All the nurses were specialists in nursing documentation; two had a long clinical experience in ICUs (N1 and N2), and one was a nursing scientist having a doctoral degree in nursing science and a position as an assistant professor in a University Department of Nursing Science (N3). The classification process was done as three separate classification tasks, i.e. each of the classes was considered separately of the others.

We measured the agreement on the content of the classes by using Cohen’s $\kappa$ [14]. Cohen’s $\kappa$ is a chance-corrected measure of agreement, which considers the classifiers equally competent to make judgments, places no restriction on the distribution of judgments over classes for each classifier, and takes into account that a certain amount of agreement is to be expected by chance. It is an appropriate measure especially in situations where there are no criteria for a correct classification, which is the case in our study.

The formulas for $\kappa$ and its standard error are the following [14]. If $P(A)$ denotes the proportion of times the judges agree, and $P(E)$ the proportion of times they would be expected to agree by chance, then

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)},$$

The standard error for $\kappa$ is

$$\sigma_\kappa = \sqrt{\frac{P(A)(1 - P(E))}{N(1 - P(E))^2}},$$

where $P(A)$ and $P(E)$ are as before, and $N$ is the total number of classified objects. This was used when computing the 95% confidence intervals for $\kappa$.

2.2.3. Regularized Least-Squares Algorithm for classification

Automatic classification was performed using Regularized Least-Squares Algorithm (RLS) (see, e.g. [15,16]). The algorithm is also known as Least-Squares Support Vector Machine [17]. The implementation of the algorithm we used was the one presented in Ref. [18], and it is briefly described in the following.

Let $(\{x_1, y_1\}, \ldots, (x_m, y_m))$ be the set of $m$ training examples, where $X$ is a set of bag-of-words vectors, $x_i \in X$ denotes the training data examples, and $y_i \in \mathbb{R}$ are their class labels set by a human expert. We consider the RLS algorithm as a special case of the following regularization problem known as Tikhonov regularization (for a more comprehensive introduction, we refer to Refs. [15,16]):

$$\min_{f} \sum_{i=1}^{m} l(f(x_i), y_i) + \lambda \|f\|^2_{k},$$

where $l$ is the loss function used by the learning machine, $f: X \rightarrow \mathbb{R}$ the function that maps the inputs $x \in X$ to the outputs $y \in \mathbb{R}$, $\lambda \in \mathbb{R}$ the regularization parameter and $\|\cdot\|_k$ is the norm in a reproducing kernel Hilbert space defined by a positive definite kernel function $k$. The second term in Eq. (3) is called a regularizer. The loss function used with RLS is the least-squares loss defined as

$$l(f(x_i), y_i) = (f(x_i) - y_i)^2.$$  

By the Representer Theorem [19], the minimizer of Eq. (3) with the least-squares loss function has the following form:

$$f(x) = \sum_{i=1}^{m} a_i k(x, x_i),$$

where $a_i \in \mathbb{R}$ and $k$ is the kernel function associated with the reproducing kernel Hilbert space mentioned above. As the kernel function, we use the cosine of the word feature vectors, that is,

$$k(x, x_i) = \frac{\cos(x, x_i)}{\sqrt{\cos(x, x)\cos(x, x_i)}}.$$  

2.2.4. Evaluation of the performance of the classification algorithm

The performance of the RLS algorithm was measured by comparing its output to the data classified manually by the nurses. The data were divided into a training set and a test set so that 708 out of the total of 1363 text pieces belonged in the training set, and the remaining 655 pieces in the test set. The division was done so that text pieces from one document belonged only in one of the two sets.

Nine automated classifiers were trained by using the training data labelled by the three nurses, i.e. one classifier for each nurse-class pair. Because of the perceived differences in the nurses’ manual classifications, we did not combine the nurses’ responses into a single data set, but considered them separately. This enabled us to study if the performance of the algorithm was affected by the different opinions of the nurses, i.e. if some manual classifications were harder for the computer to learn than the others were.

The performance of the algorithm was tested in two ways. Firstly, we tested how well the algorithm was able to capture classification patterns of one nurse by comparing the output of the algorithm with the test data manually classified
by the same nurse who also had classified the training data. Secondly, we tested how well the algorithm was able to generalize by comparing the output of the algorithm with the test data classified by other nurses than the one, with whose data the algorithm was trained. This resulted in six comparisons in each of the three classes.

The performance of the algorithm was measured with respect to the test data by using the area under ROC curve (AUC) [20], which corresponds to the probability that given a randomly chosen positive example and a randomly chosen negative example, the classifier will correctly determine which is which. Data were pre-processed with the Snowball stemmer for Finnish [21] in order to reduce different inflection forms of the words, and all the statistical analyses were performed with SPSS 11.0 for Windows.

### Table 1 – The values of Cohen’s κ and the respective 95% confidence intervals (CI) for the agreement between the three nurses

<table>
<thead>
<tr>
<th></th>
<th>Breathing</th>
<th>Blood Circulation</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>κ</td>
<td>95% CI</td>
<td>κ</td>
<td>95% CI</td>
</tr>
<tr>
<td>N1–N2</td>
<td>0.73</td>
<td>(0.68–0.78)</td>
<td>0.89</td>
</tr>
<tr>
<td>N1–N3</td>
<td>0.67</td>
<td>(0.62–0.72)</td>
<td>0.81</td>
</tr>
<tr>
<td>N2–N3</td>
<td>0.85</td>
<td>(0.82–0.89)</td>
<td>0.87</td>
</tr>
</tbody>
</table>

### Table 2 – The values of AUC and the respective 95% confidence intervals (CI) for the automatic classifiers

<table>
<thead>
<tr>
<th></th>
<th>Breathing</th>
<th>Blood Circulation</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>95% CI</td>
<td>AUC</td>
<td>95% CI</td>
</tr>
<tr>
<td>N1</td>
<td>0.86</td>
<td>(0.82–0.90)</td>
<td>0.89</td>
</tr>
<tr>
<td>N2</td>
<td>0.88</td>
<td>(0.85–0.91)</td>
<td>0.93</td>
</tr>
<tr>
<td>N3</td>
<td>0.87</td>
<td>(0.84–0.91)</td>
<td>0.91</td>
</tr>
</tbody>
</table>

### Table 3 – The values of AUC for testing the automatic classifiers against manual classifications of all the nurses

<table>
<thead>
<tr>
<th></th>
<th>Breathing</th>
<th>Blood Circulation</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>0.86</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td>N2</td>
<td>0.83</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>N3</td>
<td>0.84</td>
<td>0.88</td>
<td>0.87</td>
</tr>
</tbody>
</table>

### 3. Results

#### 3.1. Nurses’ agreement on the content of the classes

The amount of data the nurses included in the classes Breathing, Blood Circulation and Pain was, respectively, around 20, 15 and 6% of the 1363 text pieces. This reflects the extensive content of the narrative documentation, and illustrates the difficulty of finding relevant information from large amounts of text.

The comparisons between the nurses showed that the text pieces describing blood circulation were selected quite similarly (κ > 0.80 in each comparison), whereas there were more differences in selecting text pieces related to Pain or Breathing (Table 1). The range of the values of κ was the largest, from 0.67 to 0.85, in the class Breathing. In addition, given the classes Pain and Blood Circulation, the two clinical nurses N1 and N2 were the most unanimous in the classification, whereas in the class Breathing, the clinical nurse N2 and the nursing science researcher N3 were the most unanimous.

#### 3.2. Learning ability of the RLS algorithm with respect to the data classified by the nurses

The learning ability of the RLS algorithm was tested in two ways. Firstly, the algorithm was tested against the test data classified by the same nurse whose data was used when training the algorithm. The results showed that the algorithm was able to learn the classification patterns from the training set (Table 2). The values of AUC were in general around 0.85. The highest values, from 0.89 to 0.93, were achieved for the classification of blood circulation statements, whereas the lowest values, from 0.71 to 0.81 were measured for the class Pain. Except the class Pain, the performance of the algorithm was on a similar level in the classes independently of the nurse whose classification was used to train the algorithm, also in the class Breathing, in which the disagreement between the nurses was the highest. In the class Pain, the performance of the algorithm trained with the data classified by the nurse N2 was better than that of those trained with the data classified by the other two nurses.
Secondly, we tested the classifiers against the test data classified by other nurses than the one whose data was used when training the algorithm. The values of AUC were calculated for six comparisons in each of the three classes (Table 3).

The results showed that the differences in the AUCs were very small, especially in the class Blood Circulation. This means that given a manually established reference classification, the two RLS classifiers trained with other data than that of the given nurse performed as well as the classifier trained by using the classification of the given nurse. In the class Pain, the values of AUC decreased a little, except when testing against data classified by nurse N3; in this case all the classifiers performed equally well. In the class Breathing, there were larger differences in the values of AUC, and the classifiers were not able to generalize as well as in the other two classes. We also calculated the average decrease in the values of AUC compared with the situation in which both the training and the testing data were classified by the same nurse. The average decrease in the values was 0.06 in the class Breathing, 0.01 in the class Pain, and 0.00 in the class Blood Circulation.

4. Discussion

We have assessed the agreement of three nurses on the content of the classes Breathing, Blood Circulation and Pain by using Cohen’s $\kappa$, and the ability of the RLS machine-learning algorithm to learn the classification patterns of the nurses by using AUC as the performance measure. On average, the values of $\kappa$ were around 0.8. According to the obtained AUC values, the RLS algorithm performed on an acceptable level; the values were generally around 0.85, and the decreases in these were rather small when using test data classified by another nurse than the one who classified the training data.

We measured rather high values of Cohen’s $\kappa$ in the class Blood Circulation, whereas the values were somewhat lower in the other two classes. Some suggestive limits about how to interpret the observed values have been discussed within both computational linguistics [22] and medical informatics [23], concluding that the interpretation of the values depends heavily on the task, categories and the purpose of the measurement, and thus no universal guidelines can be given. In addition, the skewed distributions of categories may affect $\kappa$ [24]. In our case, all the values of $\kappa$ were above 0.67, and the majority of them were above 0.8, which indicates quite good agreement, but because of the differences in the values in the classes Breathing and Pain in particular, we had a closer look at the disagreement cases.

We analyzed the disagreement cases, and found out that the effects of the subjective considerations on important information were the most visible in the classes Breathing and Pain. For example, given the class Breathing, nurses disagreed on whether or not the statements related to the mucus in patient’s lungs should be included in the class, and given the class Pain, the disagreements were mainly due to the differences in the values in the classes Breathing and Pain. In the other two classes, the performance of the algorithm was on a similar level independently of the nurse whose manual classification was used in training, i.e. none of the classification patterns of the nurses was harder for the machine to learn than the others were.

The results also showed that in the class Blood Circulation, there were no decreases in the values of AUC when using the test data classified by another nurse than the one whose classification the algorithm was trained.
in the values of AUC in the classes Breathing and Pain reflect the nurses’ agreement on the content of these classes; in the class Blood Circulation, the values of $x$ were higher than in the classes Breathing and Pain. However, we can still conclude that based on these results, the algorithm was able to generalize substantially well.

Although our results showed that the RLS algorithm performed on an acceptable level, improvements could be gained, e.g. by using more training data, and by increasing the pre-processing of the data. In the current study, we used a stemmer to reduce different inflection forms of the words, but because Finnish is a highly inflectional language, techniques that find the real base form instead of just stemming could make the data less sparse and increase the performance of the algorithm. Another topic of further research is the pre-processing of the long sentences, which here was done manually by one researcher, and is thus a possible source of bias. Further research is also needed to assess the performance of the algorithm on other classes than the three used in this study.

In the forthcoming studies, we will test the performance using cross-validation, and thus the possible dependencies in the data set need to be taken into account. Because of the dependencies, the data set might be clustered, meaning that data points forming a cluster are more similar to each other than to other data points. With biomedical texts, it has been shown that if the data set is clustered, the traditional leave one out cross-validation may overestimate the performance of the automatic application [18]. In our case, it is possible that the text pieces extracted from one patient’s record are more similar to each other than to the text pieces extracted from another patients’ documents. In the future, we will explore this, and intend to use the leave cluster out cross-validation proposed in Ref. [18].

5. Conclusions

Our results show that the nurses seem to have somewhat different opinions on how to divide information in free text notes into different classes. We used classes Breathing, Blood Circulation and Pain, and the agreement between the nurses was highest in the class Blood Circulation, and lowest in the class Breathing. There are several possible ways to handle the disagreement cases when designing an automated classifier. The RLS algorithm was able to learn classification patterns from the data classified by the nurses, and its performance on unseen material was on an acceptable level. Based on our results, we conclude that the free text in nursing documentation can be automatically classified, and consider this as a possible way to develop electronic patient records.

Acknowledgments

We gratefully acknowledge the financial support of Tekes (grant number 40435/05) and the Academy of Finland (grant number 104639). We wish to thank Heljä Lundgrén-Laine and Filip Ginter for their contributions.

Summary points

Knowledge before this study:

- Compared to paper based patient records, electronic patient records make more data available on each patient and offer new possibilities to utilize the gathered data.
- Nursing narratives are an important part of patient documentation, but the possibilities to utilize them in the direct care process are limited due to the lack of proper tools.
- Classification of patient record narratives according to their content has been successfully applied in the medical domain. However, more research about the automated processing of nursing narratives is needed.

This study has added to the body of knowledge that:

- Nurses seem to have somewhat different opinions on how to divide information in free text notes into different classes. We used classes Breathing, Blood Circulation and Pain, and the agreement was highest in the class Blood Circulation, and lowest in the class Breathing. The disagreements should be taken into account when designing an automated classifier.
- The Regularized Least-Squares Algorithm was able to learn classification patterns from the data classified by the nurses, and its performance on unseen material was on an acceptable level.
- The free text in nursing documentation can be automatically classified and this can offer a way to develop electronic patient records.

References


