Hanna Suominen | Tapio Pahikkala | Marketta Hiissa | Tuija Lehtikunnas | Barbro Back | Helena Karsten | Sanna Salanterä | Tapio Salakoski

Relevance Ranking of Intensive Care Nursing Narratives

Turku Centre for Computer Science

TUCS Technical Report
No 740, January 2006

Relevance Ranking of Intensive Care Nursing Narratives

Hanna Suominen
University of Turku, Department of Information Technology and TUCS
hanna.suominen@utu.fi
Tapio Pahikkala
University of Turku, Department of Information Technology and TUCS
tapio.pahikkala@utu.fi
Marketta Hiissa
Åbo Akademi University, Institute for Advanced Management Systems Research and TUCS
marketta.hiissa@abo.fi
Tuija Lehtikunnas
University of Turku, Department of Nursing Science and Turku University Hospital
tuija.lehtikunnas@tyks.fi
Barbro Back
Åbo Akademi University, Institute for Advanced Management Systems Research and TUCS
barbro.back@abo.fi
Helena Karsten
University of Turku, Department of Information Technology and TUCS
eija.karsten@utu.fi
Sanna Salanterä
University of Turku, Department of Nursing Science
sanna.salantera@utu.fi
Tapio Salakoski
University of Turku, Department of Information Technology and TUCS
tapio.salakoski@utu.fi

TUCS Technical Report

No 740, January 2006

Abstract

Current computer-based patient records provide many capabilities to assist nurses' work in intensive care units, but the possibilities to utilize existing free-text documentation are limited without appropriate tools. To ease this limitation, we present an adaptation of the Regularized Least-Squares (RLS) algorithm for ranking pieces of nursing notes with respect to their relevance to breathing, blood circulation, and pain. We assessed the ranking results by using Kendall's τ_b as a measure of association between the output of the RLS algorithm and the desired ranking. The values of τ_b were 0.62, 0.69, and 0.44 for breathing, blood circulation, and pain, respectively. These values indicate that a machine learning approach can successfully be used to rank nursing notes, and encourage further research on the use of ranking techniques when developing intelligent tools for the utilization of recorded nursing narratives.

Keywords: Computerized Patient Records, Natural Language Processing, Nursing Documentation, Ranking

TUCS Laboratory
Bioinformatics Laboratory
Data Mining and Knowledge Management Laboratory
Health and Medical Informatics Institute

1 Introduction

Current computer-based patient records offer many possibilities to support nurses' work in intensive care units (ICUs). They have eased the way in which data are recorded, and they also provide new ways to utilize the recorded data. Numerical data are recorded directly from, e.g., monitoring and respiration devices, and the data are used to automatically generate different kinds of curves and diagrams. Much data are recorded also as free-text notes, but without appropriate tools, the utilization of these narratives is limited.

A possible solution to enhance using the information saved as narratives is to develop tools that process them automatically and provide nurses with targeted and relevant information, e.g., when they need to build an overview about issues related to the patient's blood circulation. To serve as a basis for these kinds of tools, we need the classification of the narratives according to their content. Furthermore, if the classification results are ranked so that each text piece is assigned a relevance score reflecting the strength of the relation to the given class, the weaker statements could be used, e.g., to focus the nurse's attention to some particular issue in order to examine it more carefully, and the stronger statements could be used, e.g., to summarize the most essential information in the record.

In health care, classification has recently been applied, e.g., to categorize texts such as chief complaint notes, diagnostic statements, and injury narratives into different kinds of syndromic, illness and cause-of-injury categories [1, 2, 3, 4, 5]. In most of these studies, the main goal was to serve the needs of different kinds of surveillance tasks. Our approach differs from these in two ways. Firstly, the aim is to develop tools that assist nurses in the direct caring process, and secondly, instead of building a binary classifier, we apply relevance ranking with respect to the given class, i.e., the more strongly the statement relates to the given class, the higher rank it gets.

In this paper, we present an adaptation of the Regularized Least-Squares (RLS) algorithm to ranking pieces of nursing notes according to their relevance to breathing, blood circulation, and pain. We chose to use the RLS algorithm, also known as the Least-Squares Support Vector Machine [6], because the Support Vector Machine approach has recently shown encouraging results, e.g., in improving the retrieval performance of WWW search engines [7], and parse ranking on a set of biomedical sentences [8].

The rest of this paper is organized as follows: Section 2 describes our data, and methods we used. Results are presented and discussed in Section 3. Finally, conclusions and future research are covered in Section 4.

2 Materials and Methods

In Section 2.1, we introduce the data and the manually performed classifications upon which the ranking system was based. Section 2.2 describes the algorithm for the automated ranking.

2.1 Materials

As data we used anonymized Finnish nursing narratives gathered with proper permissions from 16 ICUs in the spring of 2001. In total, we had 43 copies of patient records with nursing notes for one day and night. The documents contained both very long sentences describing different issues, and very short ones with the focus only on one issue. In order to standardize their style, 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.

In ranking the text pieces, we considered classes *Breathing*, *Blood Circulation*, and *Pain*. Breathing and Blood circulation were chosen because the emphasis in intensive care nursing is on monitoring, assessment and maintenance of vital functions [9]. Pain was selected for two reasons. Firstly, we think that pain assessment in critically ill patients could be supported by automatically extracting pieces of related text since, on the one hand, nurses must often perform this task relying only on implicit behavioral and physiological indicators due to the patient's inability to communicate [10], and, on the other hand, many potential pain indicators are documented in ICU nursing notes [11]. Secondly, due to the nature of pain assessment generally, and especially in critically ill patients, we assume that pain is documented differently from breathing and blood circulation.

In order to achieve the ranking that can be used in the training of the RLS algorithm and in the evaluation of the results, the text pieces were labelled independently by three nurses according to classes Breathing, Blood Circulation, and Pain. The classes were considered separately, and the nurses were advised to label a phrase if they considered it to include relevant information about the given class. The agreement between the nurses was measured by using *Cohen's* κ [12], which is a chance-corrected agreement measure, and appropriate especially in situations like ours when there are no criteria for the correctness of judgments. The values of κ showed some disagreements between the nurses, in particular in the classes Breathing and Pain (Table 1) [13].

We considered the disagreements between the nurses to reflect various interpretation possibilities as well as different experiences, knowledge and expertise areas. Thus, we regarded them as a valuable resource, upon which the ranking system was based. For each phrase, we counted the number of nurses who had labelled it belonging to the given class, and as a result of this, we got ranked data, where each phrase had a rank from zero to three corresponding to the number of

Table 1: The values of Cohen's κ and respective 95 % confidence intervals (CI) for comparisons between nurses N_1, N_2, and N_3 [13].

	Breathing		Blood Circulation		Pain	
	κ	95 % CI	κ	95 % CI	κ	95 % CI
$N_1 - N_2$	0.73	(0.68 – 0.78)	0.89	(0.85 – 0.92)	0.88	(0.82 – 0.94)
$N_1 - N_3$	0.67	(0.62 – 0.72)	0.81	(0.77 – 0.86)	0.79	(0.73 – 0.86)
$N_2 - N_3$	0.85	(0.82 – 0.89)	0.87	(0.83 – 0.90)	0.76	(0.69 – 0.83)

nurses who had assigned the phrase into the class. This ranking was considered to reflect the relevance level of the expressions according to the given class. The ranked data were divided into *training set* and *test set* so that 708 pieces of text belonged to the training set, and the remaining 655 pieces to the test set. No patient record was divided between the two sets.

2.2 Methods

The automated classification was performed by using a *Regularized Least-Squares* (*RLS*) *algorithm*. Let $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, where X is a space of bag-of-words vectors, $x_i \in X$, and $y_i \in \{0, 1, 2, 3\}$, be the set of training examples. The target output value y_i is the rank of the corresponding example, set according to the labellings of the three nurses. We consider the RLS algorithm as a special case of the following regularization problem known as the Tikhonov regularization (for a more comprehensive introduction, see, e.g., [6], or [14]):

$$\min_f \sum_{i=1}^{m} l\left(f(x_i), y_i\right) + \lambda \|f\|_k^2, \tag{1}$$

where l is the loss function used by the learning machine, $f : X \to \mathbb{R}$ is a function, $\lambda \in \mathbb{R}_+$ is a regularization parameter, and $\|\cdot\|_k$ is a norm in a reproducing kernel Hilbert space defined by a positive definite kernel function k. The second term in (1) is called a *regularizer*. The loss function used with RLS is called *least-squares loss* and is defined as

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

By the Representer Theorem, the minimizer of equation (1) has the form

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

where $\alpha_i \in \mathbb{R}$ and k is the kernel function associated with the reproducing kernel Hilbert space mentioned above. The kernel function which we use is the cosine of the word feature vectors, i.e.,

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

The RLS algorithm was trained for each of the classes separately. To unify the data, we separated punctuation marks as well as special characters from words with spaces, converted all letters into lower case letters, and used *the Snowball stemming algorithm* [15] for Finnish text. In order to optimize the ranking of the training examples, we divided the output values into four separate intervals corresponding to the four labelling levels. The threshold values that determined the output intervals and the regularization parameter were selected by using a leave-one-out cross-validation on the training set with a rank correlation coefficient *Kendall's* τ_b [16] as the optimization criterion. We chose to use this measure of the association because, according to [16], it is appropriate in situations like ours when there is a considerable amount of tied items in the data.

3 Results and Discussion

We measured the association between the rankings produced by nurses and the RLS algorithm by using Kendall's τ_b. According to these values (Table 2), the strongest associations between the desired ranking and the one produced by the RLS algorithm were reached in classes Blood Circulation (τ_b=0.69) and Breathing (τ_b=0.62). In the class Pain, the value of τ_b was 0.44.

Next, a possibility to utilize the ranking is illustrated in Figure 1 by considering the case in which there is a need to quickly build an overview about issues related to blood circulation in our test set. In this case, by choosing the sensitivity level from the output of the RLS algorithm, e.g., the fifty or hundred highest ranked phrases of nursing notes could be returned. In our first experiment, when fifty phrases were returned, a great majority of them (45 phrases) belonged to Rank 3 in the desired ranking. Two phrases (*Situation stable.* and *Lift in infusion.*) were considered as irrelevant in the desired ranking, and the number of returned phrases with ranks one and two in the desired ranking were one and two, respectively. In our second experiment with hundred phrases, 55 out of 65 phrases with the rank three in the desired ranking were returned. Here, however, the number of phrases that were considered as irrelevant in the desired ranking increased to 39, and the number of returned phrases with ranks one and two in the desired ranking were one and five, respectively. In summary, this example demonstrates how ranking tools coud enhance the utilization of existing narratives.

Let us now consider the manual ranking upon which the RLS algorithm was based. In this ranking, classes Breathing, Blood Circulation, and Pain had similar general characteristics of the ranks. For example, in the class Breathing, the phrases with the rank three were related to issues such as oxygenation and the use

Table 2: The values of Kendall's τ_b and respective 95 % confidence intervals (CI), calculated with SPSS 11.0 for Windows, for rankings in classes Breathing, Blood Circulation, and Pain.

	τ_b	95% CI
Breathing	0.62	(0.56 – 0.68)
Blood Circulation	0.69	(0.61 – 0.76)
Pain	0.44	(0.30 – 0.59)

of different kinds of respiration devices, whereas the phrases with lower ranks described slime in patients' lungs, coughing, and issues related to pleural tubes. In the class Blood Circulation, the phrases with the highest rank were mostly about the progression of various measurement values, such as pulse or blood pressure. The phrases with the blood circulation ranks two or one described issues such as patients' body temperature and bluish skin shade. In the class Pain, the phrases with the rank three commonly contained keywords such as *pain* or *ache*, whereas the phrases with rank one or zero included neither the word *pain* nor its derivatives or synonyms. The phrases with the pain rank one or two covered implicit pain indicators such as the quality of sleep and reactions to nursing interventions. These examples illustrate that the phrases which all the nurses considered to contain relevant information about the particular class were the most evident notes about the class. The connection to the topic got more implicit when the relevance rank decreased. Hence, it seems that the manual ranking logically demonstrates the connection of the phrases and the given class. However, more work is required in order to obtain a ranking that reliably reflects the real relevance of the phrases with respect to the desired classes.

In all three classes, the differences between the manual and automated rankings seemed to be mostly due to the sparse data. Thus, the performance of the RLS algorithm might be enhanced by reducing the sparseness of the data through the use of pre-processing algorithms that find the base form of the word instead of its stem. For example, this kind of an algorithm may enhance the performance of the RLS algorithm by recognizing the key word *kipu* (pain) from its derivatives *kivulias* and *kipuja*, and from rare compounds, such as *kipulääke* (painkiller) and *kipukynnys* (pain threshold). Evidence of the importance of the pre-processing algorithm in the unification of the data was also provided by our preliminary experiments in which we found that the performance of the RLS algorithm was better with the stemmed data than with the original raw data.

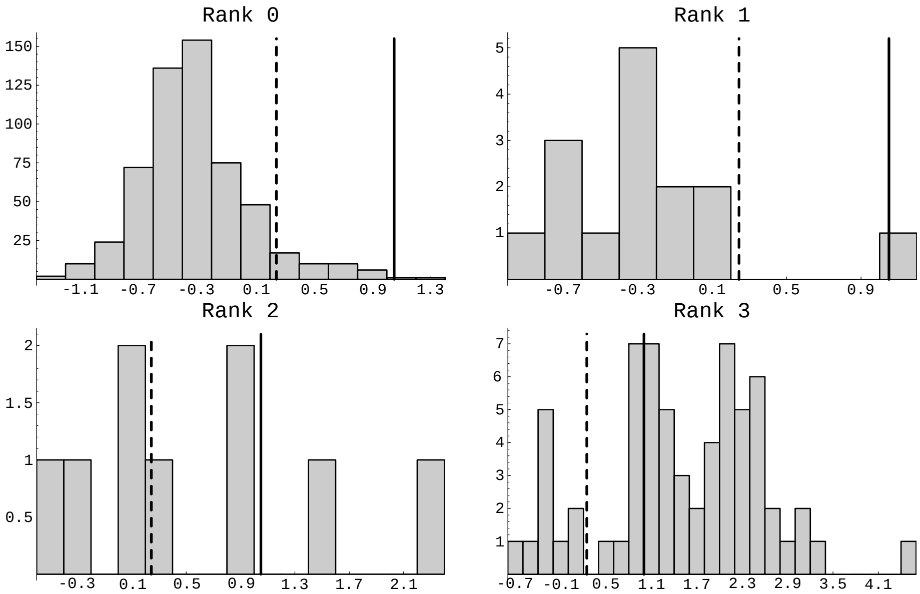

Figure 1: Histograms illustrating the association between the desired ranking and the output of the RLS algorithm for each of the four human ranks separately in the class Blood Circulation. The horizontal axis corresponds the output of the RLS algorithm in real values, and the vertical axis represents the number of phrases in the corresponding range. Vertical lines at 1.05 (solid) and at 0.24 (dashed) separate the output of the RLS algorithm so that 50 or 100, respectively, of the highest ranked phrases are on the right-hand side of the line.

The performance of the RLS algorithm could also be enhanced by using larger training set because the larger training set is more likely to contain more instances of infrequently occurring words such as names of medicines or devices and non-standard abbreviations. For example, the insufficient number of training instances related to pain can explain the weaker results in the class Pain since all the phrases including the word *headache* were ranked mistakenly to the pain rank zero instead of the correct rank three because the training set did not contain this word. Moreover, altogether only 62 out of 708 training instances belonged to higher pain ranks than Rank 0. The corresponding numbers were 136 and 144 in classes Breathing and Blood Circulation, in which the RLS algorithm learnt nurses' ranking better.

In addition to pre-processing of the data, more work is needed in order to divide long sentences into smaller pieces automatically. However, although the division we used was based on the point of view of one nursing science researcher, it can be considered quite objective because only few lines belonged to two different classes and none of text lines belonged to three classes in any of the manual classifications.

4 Conclusions and Future Research

The association between the output of the RLS algorithm and the desired ranking was somewhat weaker in the class Pain than in classes Breathing and Blood Circulation. We consider the differences in the strength of the associations to be due to the different nature of the documentation related to the classes, and anticipate this to affect the details of the ranking application.

Our results encourage further research on the use of ranking techniques when developing intelligent tools for the utilization of recorded nursing narratives. In the future, tools of this kind may enable, e.g., automated highlighting of nursing narratives according to their relevance to the particular topic. In this way, the utilization of recorded narratives can be supported without losing the original context of text pieces. Moreover, the ranking result reflecting the relevance of the text pieces with respect to the desired classes is more informative than the classification output which only separates relevant text pieces from irrelevant parts of the documentation.

Acknowledgements

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

References

[1] Chapman, W.W., Dowling, J.N., Wagner M.M.: Fever detection from free-text clinical records for biosurveillance. *Journal of Biomedical Informatics 37*: 120–7. 2004.

[2] Pakhomov, S.V., Buntrock, J.D., Chute, C.G.: Using compound codes for automatic classification of clinical diagnoses. In: Fieschi, M., Coiera, E., Li, Y.-C. *Proceedings of the 11th World Congress on Medical Informatics*: 411–5. Hilton, San Fransisco, USA, Sept 7–11, 2004,

[3] Wellman, H.M., Lehto, M.R., Sorock, G.S., Smith, G.S.: Computerized coding of injury narrative data from the National Health Interview Survey. *Accident, Analysis and Prevention 36*: 165–71. 2004.

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

[5] Pakhomov, S.V., Buntrock, J., Chute, C.G.: Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *Journal of Biomedical Informatics 38*: 145–53. 2005.

[6] Rifkin, R.M.: *Everything old is new again: A fresh look at historical approaches in machine learning*. PhD Thesis, Massachusetts Institute of Technology. 2002.

[7] Joachims, T.: Optimizing search engines using clickthrough data. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23–26, 2002. New York, NY, USA, ACM Press. 133–42. 2002.

[8] Tsivtsivadze, E., Pahikkala, T., Pyysalo, S., Boberg, J., Mylläri, A., Salakoski, T.: Regularized Least Squares for parse ranking. In A.F. Famili et al. (Eds.): *IDA 2005*, *Lecture Notes in Computer Science 3646*. Springer-Verlag Berlin Heidelberg. 464–74. 2005.

[9] Ward, N.S., Snyder, J.E., Ross, S., Haze, D., Levy, M.M.: Comparison of a commercially available clinical information system with other methods of measuring critical care outcomes data. *Journal of Critical Care 19*: 10–5. 2004.

[10] Payen, J.-F., Bru, O., Bosson, J.-L., Lagrasta, A., Novel, E., Deschaux, I., Lavagne, P., Jacquot, C.: Assessing pain in critically ill sedated patients by using a behavioral pain scale. *Critical Care Medicine 29*: 2258–63. 2001.

[11] Gélinas, C., Fortier, M., Viens, C., Fillion, L., Puntillo, K.: Pain assessment and management in critically ill intubated patients: a retrospective study. *American Journal of Critical Care 13*: 126–35. 2004.

[12] Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*: 37–46. 1960.

[13] Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., Salakoski, T.: Towards automated classification of intensive care nursing narratives. *Technical Report Number 739, Turku Centre for Computer Science (TUCS)*. 2006.

[14] Poggio, T., Smale, S: The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society 50*: 537–44. 2003.

[15] *Snowball, the Finnish stemming algorithm*. Martin Porter; c2001 and (for the Java developments) Richard Boulton; c2002 [updated 2005 May 23; cited 2006 Jan 10]. http://snowball.tartarus.org/algorithms/finnish/stemmer.html

[16] Kendall, M.G.: *Rank correlation methods*. 4th edition, Griffin, London. 1970.

TURKU

CENTRE *for*

COMPUTER

SCIENCE

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 952-12-1683-2
ISSN 1239-1891