Evaluation of a Prototype System that Automatically Assigns Subject Headings to Nursing Narratives Using Recurrent Neural Network

Hans Moen[1], **Kai Hakala**[1], **Laura-Maria Peltonen**[2,3], **Henry Suhonen**[2,3], **Petri Loukasmäki**[1], **Tapio Salakoski**[1], **Filip Ginter**[1] **and Sanna Salanterä**[2,3]
[1]Turku NLP Group, Department of Future Technologies, University of Turku, Finland
[2]Department of Nursing Science, University of Turku, Finland
[3]Turku University Hospital, Finland
{hanmoe,kahaka,lmemur,hajsuh,
peerlo,figint,tapio.salakoski,sansala}@utu.fi

Abstract

We present our initial evaluation of a prototype system designed to assist nurses in assigning subject headings to nursing narratives – written in the context of documenting patient care in hospitals. Currently nurses may need to memorize several hundred subject headings from standardized nursing terminologies when structuring and assigning the right section/subject headings to their text. Our aim is to allow nurses to write in a narrative manner without having to plan and structure the text with respect to sections and subject headings, instead the system should assist with the assignment of subject headings and restructuring afterwards. We hypothesize that this could reduce the time and effort needed for nursing documentation in hospitals. A central component of the system is a text classification model based on a long short-term memory (LSTM) recurrent neural network architecture, trained on a large data set of nursing notes. A simple Web-based interface has been implemented for user interaction. To evaluate the system, three nurses write a set of artificial nursing shift notes in a fully unstructured narrative manner, without planning for or consider the use of sections and subject headings. These are then fed to the system which assigns subject headings to each sentence and then groups them into paragraphs. Manual evaluation is conducted by a group of nurses. The results show that about 70% of the sentences are assigned to correct subject headings. The nurses believe that such a system can be of great help in making nursing documentation in hospitals easier and less time consuming. Finally, various measures and approaches for improving the system are discussed.

1 Introduction

An important task for hospital nurses is to document the administrated patient care in order to ensure care continuity. These nursing (shift) notes are typically stored in patients' electronic health records. However, documentation constitutes a relatively large portion of nurses time, up to 35%, and an average of 19% (Yee et al., 2012). Reducing the time spent on documentation will free up more time for direct patient care. As a means to make the documented text easier to navigate and process, e.g., for the purpose of planning and extracting statistics, nurses in many countries are required to perform some sort of structuring of the text they write (Saranto et al., 2014). Such structuring approaches include the use of documentation standards, classifications and standardized terminologies (Hyppönen et al., 2014). Compared to using fully unstructured free (narrative) text, certain restrictions and requirements to the documentation process are added. As an example, in Finland nurses are nowadays expected to structure the text they write by using subject headings from the Finnish Care Classification (FinCC) standard (Hoffrén et al., 2008). FinCC consist primarily of two taxonomy resources, the Finnish Classification of Nursing Diagnoses (FiCND) and the Finnish Classification of Nursing Interventions (FiCNI), and both of these have a three-level hierarchy. For example, one branch in FiCND is: "Tissue integrity" (level 1), "Chronic wound" (level 2) and "Infected wound" (level 3). Another example, a branch from FiCNI is: "Medication" (level 1), "Pharmacotherapy" (level 2) and "Pharmaceutical treatment, oral instructions" (level 3). In sum, FinCC consist of more than 500 subject headings, making it challenging and time consuming for nurses to use since they are required to memorize, use and structure the text they write according to such a large number of subject headings (Häyrinen et al., 2010).

Our goal is to assist nursing documentation by developing a system that is able to automatically, or semi-automatically, assign subject headings to

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 94–100
Brussels, Belgium, October 31, 2018.

nursing narratives according to the current care classification standard. A central component is a text classification model based on a long short-term memory (LSTM) recurrent neural network architecture (Hochreiter and Schmidhuber, 1997; Gers et al., 2000). We hypothesize that such a system has the potential to reduce the time and effort needed for documentation. It could also increase the consistency in the use of subject headings, and potentially improve the documentation quality. We see two use-cases for such a system: One is where the system assists nurses in selecting appropriate headings when they write, in a suggestive manner, e.g., per sentence or paragraph; A second use-case is where nurses are allowed to write in an unstructured narrative manner, without having to take into consideration the use of subject headings. Instead the system should assign subject headings afterwards and restructure the text under the various subject headings when such a representation is needed. In the presented experiment we focus on the second use-case, where we evaluate the performance of a prototype system developed for this purpose.

2 Related Work

Natural language text is among the most complex data types commonly used for storing and managing information. Thanks to continuous advancements in the field of natural language processing (NLP), computers are becoming capable of performing increasingly complex tasks on this type of data.

Denny et al. (2009) present an algorithm called "SecTag" for detecting section headers in clinical notes based on the free text. More precisely, they focus on history and physical examination documents where the goal is to identify and normalize section headers as well as to detect section boundaries, evaluated with 29 section headers to choose from. For this they use various NLP techniques including word recognition, terminology-based rules, and naive Bayesian classifier. Li et al. (2010) present a system that categorizes sections in clinical notes into one of 15 pre-defined section labels. They use a Hidden Markov model which expects as input clinical notes that have already been split into sections. In Haug et al. (2014) the goal is to develop a "Clinical Section Labeler" which assigns standardized topics to the sections found in clinical notes. These topics, 28 in total, are here seen as separate from the section headings used by the clinicians when writing, thus the section headings are considered as input to the classifier along with the free text. As classifiers they use two variations of Bayesian networks.

Deep learning methods based on artificial neural networks (ANNs) are currently representing state of the art in many NLP tasks (Zhang et al., 2015; Tang et al., 2015), including text classification, relation extraction and translation. In the presented experiment/prototype system we use the popular long short-term memory (LSTM) recurrent neural network architecture (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) for conducting the text classification. In the data set used here there are 676 unique headings to choose from by the classifier.

3 Methods

3.1 User Interface

The prototype system is implemented in Python with a simple Web interface using the Flask framework (Grinberg, 2018). The interface allows users to upload a text document (i.e. nursing narrative), as shown in Figure 1. When pressing the *create headings button*, the system first splits the text into sentences and then performs word-level tokenization. Each sentence is then fed to the pre-trained text classification model (described below) which assigns subject headings the one subject heading with the highest confidence score according to the classifier. Based on their assigned subject headings, sentences are grouped into paragraphs – one paragraph per unique subject heading. Figure 2 shows a translated example of how a nursing note without subject headings (upper) is converted into paragraphs with assigned subject headings (lower) using the system. Although not utilized in the experiment presented here, the interface also allows the user to move sentences between paragraphs, edit existing subject headings and add new subject headings/paragraphs. In addition, when holding the mouse cursor over a sentence the system shows its top 10 subject heading suggestions according to the classifier. These features provide the user with ways to quickly correct the initial subject heading assignments conducted by the system.

3.2 Text Classification

A central component of the system is the text classification model. The classification task is ap-

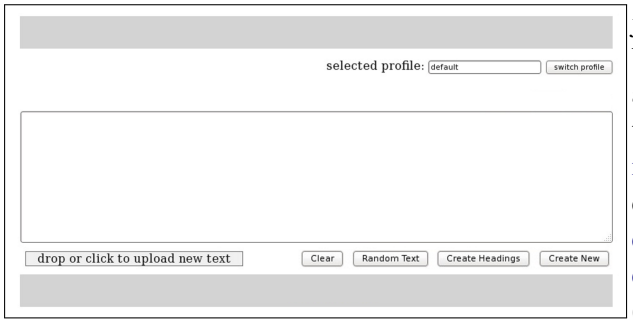

Figure 1: Prototype Web interface.

Oxynorm 10 mg p.o. for abdominal pain when needed to relieve pain. Eaten breakfast. NaCl 0.9 1———-1 cannula removed. Reads news and watches TV in recreation room after breakfast. Feeling well and pain free at the time, the oxynorm administered in the morning helped. CRP decreased now 63, leuc 7.4, also in decline. Eaten lunch. Sister visits after lunch. No need for a sickness certificate. Wound treatment instructions and pain prescriptions given. Has permission to go home in the evening, sister comes to pick up at some point. Left for home at 18.30.

PAIN
Oxynorm 10mg p.o. for abdominal pain when needed to relieve pain.

NUTRITION
Eaten breakfast.
Eaten lunch.

FLUID THERAPY
NaCl 0,9 1———-1 cannula removed.

CURRENT HEALTH AND FUNCTIONALITY
Reads news and watches TV in recreation room after breakfast.
Feeling well and pain free at the time, the oxynorm administered in the morning helped.
Sister visits after lunch.
Left for home at 18.30.

DOCTORS VISIT
CRP decreased now 63, leuc 7.4, also in decline.
No need for a sickness certificate.
Has permission to go home in the evening, sister comes to pick up at some point.

EDUCATION OF RELATIVES
Wound treatment instructions and pain prescriptions given.

Figure 2: An example showing how a nursing note written in a purely narrative manner (upper) is assigned headings and structured using the system (lower). This has been translated from Finnish to English.

proached as a multiclass classification task, where each sentence is assumed to have one correct subject heading (i.e. class/label). There exist a number of different methods and tools that are suitable for this type of text classification, including the already mentioned LSTM networks (Hochreiter and Schmidhuber, 1997; Gers et al., 2000), convolutional neural networks (CNNs) (LeCun et al., 1998), Random Forest classifiers (Liaw et al., 2002) and support vector machine classifiers (SVM) (Joachims, 1999). However, the focus of this study is not to find the optimal text classification method and parameter settings for this task. This has been the focus of a previous study (under review), where a range of different state-of-the-art and baseline text classification methods are tested and compared. The mentioned study indicated that a bidirectional version of LSTM networks performs best when compared to other classification methods/models, including CNN, SVM and Random Forest. A LSTM network is designed to process sequential data in that it makes its final classification decision after having iteratively observed each element in a sequence, where the order of the elements matters. In our case, a sequence is a list of words belonging to a sentence. This ability to utilize word ordering and to detect long distance word relations in the input sentences is a strength of LSTM networks compared to other text classification approaches relying on bag of word features. In the bidirectional version of LSTM that we use, a sentence is read from both left to right and right to left. This network has been trained on the training set described below. We use the Python-based Keras deep learning library (Chollet et al., 2015) with Theano tensor manipulation library (Bastien et al., 2012) as backend engine.

3.3 Training Data

The data set used for training the classifier is a collection of approximately 0.5 million patients' nursing notes extracted from a hospital in Finland. Ethical approval for using the data was obtained from the hospital district's ethics committee and research approval was obtained from the medical director of the hospital district. The selection criteria were patients with any type of heart-related problem in the period 2005 to 2009. This includes nursing notes from all units in the hospital visited during their hospital stay. The data is collected during a transition period between an older care classification standard and the mentioned FinCC standard, thus only a subset of the headings found

CANNULA CARE
Taken care of the cannula himself. Bandage contains stringy colourful mucus. NaCl cleaning + change of bandages.

taken care of the cannula himself .	*cannula_care*
bandage contains stringy colourful mucus .	*cannula_care*
nacl cleaning + change of bandages .	*cannula_care*

Figure 3: An example showing how a paragraph (upper) is converted into a set of sentence-level training examples (lower). This has been translated from Finnish to English.

there are from FinCC. We only use sentences occurring in a paragraph with a subject heading, which amounts to approximately 5.5 million sentences, 133,890 unique tokens and approximately 38.5 million tokens in total. The average sentence length is 7 tokens and the average number of sentences per paragraph is 2.1. To reduce the number of unique subject headings and to ensure that each included subject heading has a fair number of training examples, we apply a lower frequency threshold of 100. This result in 676 unique subject headings, where their frequency count range from 100 to 222,984, with an average of 4,896. We convert the data into training examples by splitting each paragraph into sentences, each representing a training example with input (X) being the sentence and the output (y) being the associated subject heading of the paragraph. See Table 3 for an example. This enables classification on sentence level, which further allows restructuring and grouping of sentences that are classified as having the same or similar headings. The data set was split into training (60%), development (20%) and test (20%) sets.

Although not the focus of this paper, we report the performance of the bidirectional LSTM classifier when used to predict subject headings for the test set, as a comparison to the experiment presented below. Performance is calculated as recall at N (R@N), which is the average of how many times the correct subject heading is found among the top N suggested subject headings by the system. R@1 is here equal to the classifier's accuracy score on the test set. These results are presented in Table 1. We refer to this evaluation as an automatic evaluation since no (additional) manual evaluation is required.

Measure	**Score**
R@1 / Accuracy	54.35%
R@10	89.54%

Table 1: The classifiers performance on the test set. R@N is recall at N, reflecting the average of how many times the correct subject heading is found among the top N retrieved ones, over all sentences, in the test set. R@1 is equal to accuracy.

4 Experiment

The main objective of the experiment is to assess how well the described system is able to assign relevant subject headings to nursing notes that are written in a narrative manner, without using or considering subject headings. A secondary objective is to report on feedback from nurses concerning the potential use of such a system in a clinical setting.

The nursing notes that we have in the existing data set are all planned, written and structured according to the ruling documentation standard – where the text is split into sections labeled with subject headings. Thus, to acquire relevant nursing notes for the evaluation – nursing notes written in a way where the authors does not plan for or consider the use of sections and subject headings – we asked three domain experts with nursing background to write a couple of notes each in this way based on made up artificial patients. This resulted in a total of 20 nursing notes. These were then presented to the system, one by one, which classified and assigned subject headings on sentence level before grouping sentences under each heading. The results were stored in a spreadsheet for evaluation, containing a short description of the patient case, the original nursing note and the version with assigned subject headings on sentence level. See Figure 2 for an example of one of the nursing narratives/notes used in the evaluation, both without and with the assigned headings and restructuring conducted by the system.

Next, two domain experts (hereby referred to as evaluators) were given the task of assessing how well the system performed. For this the evaluators were (a) instructed to use a four class scale when manually assessing each sentence with respect to their assigned headings, and (b) asked to answer the open ended question "what do you think about the current performance and functionality of the system and its potential use in a clinical setting?".

Class	Count	Percentage
1 / Accuracy$_{min}$	311	68.05%
2	93	20.35%
3	48	10.50%
4	5	1.10%
1 + 2 / Accuracy$_{max}$	404	88.40%

Table 2: Average results from the manual evaluation. Class description: 1 - Correct heading. 2 - Maybe correct heading. 3 - Wrong heading. 4 - Unable to assess.

The four classes are as follows:

1 - Correct heading (it correctly describes the content of the sentence)

2 - Maybe correct heading

3 - Wrong heading

4 - Unable to assess

The proportion of sentences assigned to Class 1 is equal to the accuracy$_{min}$ score of the system for this task, while the sum of Class 1 and 2 can be considered as the accuracy$_{max}$ score. So the actual accuracy score would be somewhere between accuracy$_{min}$ and accuracy$_{max}$.

5 Results

Initially the two evaluators disagreed in their assessments of 30.45% of the sentences. To reach a common consensus, the two evaluators discussed these cases together with a third domain expert. The results from the manual evaluation (consensus) are presented as average counts and percentages for each class in Table 2.

The percentage of correctly classified sentences in the manual evaluation experiment is 68.05% (Table 2). However, the actual accuracy score of the system can be assumed to be somewhere between 68.05% (accuracy$_{min}$) and 88.40% (accuracy$_{max}$). This is roughly 13% to 34% points up from the R@1/accuracy score resulting from the automatic evaluation in Table 1. When the system is allowed to suggest 10 headings, R@10, the correct heading is found among these for about 90% of the sentences in the test set. I.e. at least one of the suggested 10 headings for a sentence has been considered correct for about 90% of the test set sentences in the manual evaluation.

The evaluators reported that they were generally satisfied with the performance of the system. They think that such a system/functionality could be very useful to have as an integrated part of a hospital information system/electronic health record system, and could reduce the time and effort required to perform the documentation. They also think that it has the potential to increase the quality of documentation by supporting the correct use of such standardized terminologies. The evaluators reported that the system showed a tendency to assign subject headings with a high level of specificity, and sometimes even too specific than what would be practical. For example, for two or more sentences describing different aspects of pain management in the same nursing note, such as treatment and medication, the system would in some cases assign these to different subject headings, and/or headings of different level of specificity/abstraction. Another observation was that the system had sometimes difficulties in correctly classifying sentences that covers multiple subjects.

6 Discussion

One obvious observation is that there is a relatively large gap between the scores resulting from the conducted manual evaluation (68.05% $\leq$ accuracy $\leq$ 88.40%, Table 2) and the automatic evaluation scores (accuracy = 54.35%, Table 1). We believe that this is caused by primarily two underlying problems: First, the data set spans two different documentation standards (as described in Section 3), which could be somewhat confusing to the classifier. Second, the nurses do not necessarily always use the correct subject headings when they write. Thus it is likely, in particular for this type of automatic evaluation, that higher scores will be achieved when the classifier is trained and evaluated on a data set consisting of only one documentation standard. When looking at the R@10 scores (Table 1), the system suggests the correct heading for about 90% of the sentences in the test set. However, it is likely that the same problem of "classification standard confusion" negatively influences this score too. For a use-case where, let us say, the system suggests 10 headings per sentence to the user when he/she is writing the nursing notes, this would mean that there is a very high probability ($\geq$ 90%) of finding a suitable/correct subject heading among the suggested ones.

Based on their observations, the evaluators found the system to sometimes assign subject headings with an artificial detail level. One way to

deal with this would be to allow the users to preselect the level in the hierarchy of the documentation standard that the system should aim for when assigning subject headings. In addition, since a unit in the hospital would typically not use all the headings in the documentation standard, it should be possible to limit the headings that the system can choose from for different units.

To further improve the performance of the system there are several, possibly complementary, approaches that could be explored. One approach is to allow the user to manually correct the initial classifications done by the system, e.g. by moving sentences to their correct subject headings, and allowing the user to add and remove subject headings at will. Additionally, this type of manual corrections could be used to further improve the system/classifier. A possibly complementary approach could be to apply some form of classification heuristic and/or feedback based on the confidence scores produced by the classifier. For example, when classifying a sentence, if the classifier shows very similar confidence scores for the top suggested subject headings, and if a subject heading used in the same or a previous nursing note, from the same patient and care episode, is among these, one could have the system select this one. Another example, if the classifier does not show a clear preference for a single subject heading when classifying a sentence, this could be communicated to the user. Some type of clustering of subject headings that are very similar (in terms of form and/or meaning) within a single nursing note could also be tried. It would also make sense to exploit the taxonomic hierarchy underlying the nursing documentation standard, e.g. during training and/or prediction as well as in the grouping of sentences and possibly for merging some of the assigned subject headings. Another approach would be to try using a more balanced data set for training the classifier – balanced in terms of label/subject heading frequencies. The use of class weighting when training the classifier could also be tried. With enough training data, it could also be an idea to train a separate classifier per hospital unit. Further performance gains could be achieved by also training a classifier on the level of paragraphs as a supplement to the sentence-level classification.

Although the focus of this work has been on assisting nursing documentation, other professions use subject headings in a similar fashion when they write. One example is physicians and the notes they write in relation to diagnosis and treatment of patients. Thus we assume that the same type of classification-based system could be useful to other professions too.

7 Conclusions and Future Work

The presented prototype system for automated assignment of subject headings to nursing notes is shown to perform well based on the reported experiment. It achieves a classification accuracy somewhere between 68.05% (accuracy$_{min}$) and 88.40% (accuracy$_{max}$). The domain experts evaluating the system reported that they believe such a system could save both time and effort when it comes to writing nursing shift notes in hospitals. We argue that future improvements of the system's classification performance could be gained through user feedback or by applying some heuristic based on its confidence scores. In the presented experiment we have the classification system learn to classify text on the level of sentences. As future work we are also considering exploring paragraph-level classification for this task, primarily as a supplement to sentence-level classification. Since there are other professions who use subject headings in a similar way as nurses when they document, we believe that a similar system could also be useful in other domains, for other professions.

As future work we aim to test this system/classifier on a larger scale, where it will also be evaluated when used in the initial writing of nursing notes, by suggesting N subject headings to the user for each sentence being written. We will also strive to acquire a data set containing only one documentation standard – the one currently being used in the targeted hospital district. Then the following step would be clinical testing and assessment of the impact of such a system (extrinsic evaluation).

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: New features and speed improvements. *arXiv preprint arXiv:1211.5590 (2012).*

François Chollet et al. 2015. Keras. https://keras.io.

Joshua C. Denny, Anderson Spickard, III, Kevin B. Johnson, Neeraja B. Peterson, Josh F. Peterson, and Randolph A. Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471.

Miguel Grinberg. 2018. *Flask web development: developing web applications with Python*. " O'Reilly Media, Inc.".

Peter J Haug, Xinzi Wu, Jeffery P Ferraro, Guergana K Savova, Stanley M Huff, and Christopher G Chute. 2014. Developing a section labeler for clinical documents. In *AMIA Annual Symposium Proceedings*, volume 2014, page 636. American Medical Informatics Association.

Kristiina Häyrinen, Johanna Lammintakanen, and Kaija Saranto. 2010. Evaluation of electronic nursing documentation – Nursing process model and standardized terminologies as keys to visible and transparent nursing. *International Journal of Medical Informatics*, 79(8):554–564.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Päivi Hoffrén, Kirsi Leivonen, and Merja Miettinen. 2008. Nursing standardized documentation in kuopio university hospital. *Studies in Health Technology and Informatics*, 146:776–777.

Hannele Hyppönen, Kaija Saranto, Riikka Vuokko, Päivi Mäkelä-Bengs, Persephone Doupi, Minna Lindqvist, and Marjukka Mkelä. 2014. Impacts of structuring the electronic health record: A systematic review protocol and results of previous reviews. *International Journal of Medical Informatics*, 83(3):159–169.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 744–750, New York, NY, USA. ACM.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news*, 2(3):18–22.

Kaija Saranto, Ulla-Mari Kinnunen, Eija Kivekäs, Anna-Mari Lappalainen, Pia Liljamo, Elina Rajalahti, and Hannele Hyppönen. 2014. Impacts of structuring nursing records: a systematic review. *Scandinavian Journal of Caring Sciences*, 28(4):629–647.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432.

Tracy Yee, Jack Needleman, Marjorie Pearson, Patricia Parkerton, Melissa Parkerton, and Joelle Wolstein. 2012. The influence of integrated electronic medical records and computerized nursing notes on nurses time spent in documentation. *Computers Informatics Nursing*, 30(6):287–292.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.