Supporting Communication and Decision Making in Finnish Intensive Care with Language Technology

Hanna J. Suominen[1*] and Tapio I. Salakoski[2]
[1]Canberra Research Laboratory, NICTA and College of Engineering and Computer Science, Australian National University, Canberra, Australia
[2]Turku Centre for Computer Science (TUCS) and Department of Information Technology, University of Turku, Turku, Finland

ABSTRACT
A fluent flow of health information is critical for health communication and decision making. However, the flow is fragmented by the large amount of textual records and their specific jargon. This creates risks for both patient safety and cost-effective health services. Language technology for the automated processing of textual health records is emerging. In this paper, we describe method development for building topical overviews in Finnish intensive care. Our topical search methods are based on supervised multi-label classification and regression, as well as supervised and unsupervised multi-class classification. Our linguistic analysis methods are based on rule-based and statistical parsing, as well as tailoring of a commercial morphological analyser. According to our experimental results, the supervised methods generalise for multiple topics and human annotators, and the unsupervised method enables an ad hoc information search. Tailored linguistic analysis improves performance in the experiments and, in addition, improves text comprehensibility for health professionals and laypeople. In conclusion, the performance of our methods is promising for real-life applications.

Keywords: computer-assisted decision making, electronic health records, intensive care, medical informatics applications, natural language processing, nursing

1. INTRODUCTION
1.1. Background
A fluent flow of health information is critical in decision making – both for health professionals in their work and for laypeople in managing their health in various life situations (Figure 1). *Flow of information* is defined as links, channels, and contact, or the flow of communication to pertinent people or groups in the organisation [1]. *Health information* refers to any information that

a) is created or received by a healthcare provider, health plan, public health authority, employer, life insurer, school or university, or healthcare clearing-house, and
b) relates to the past, present, or future physical or mental health or condition of an individual, the provision of healthcare to an individual, or the past, present, or future payment for the provision of healthcare to an individual [2].

[*]Corresponding author: NICTA, Locked Bag 8001, Canberra ACT 2601, Australia; hanna.suominen@nicta.com.au

The primary purpose of this information is to serve patient care and its continuity as an intermediary communication within a multi-professional team of health practitioners as well as between health professionals and laypeople [3]. For ensuring the fluent flow of health information, health professionals are obligated by law [4]

a) to document this information in health records correctly, clearly, and understandably,
b) to document all necessary and sufficient information needed for decision making related to organising, planning, performing, and controlling good quality healthcare, and
c) to use generally known and widely accepted terminology and abbreviations.

These definitions not only relate the concepts of decision making, communication, and flow of information to another, but also create the theoretical underpinning of this paper: getting the right health information to the right people in the right format at the right time. The fluent flow of information is a prerequisite for precise healthcare decision making and this includes communication within and between laypeople and the multi-professional team of health practitioners. With the right information and time, we refer to the different needs in decision making and communication. With the right people and format, we refer to the varying needs of different users.

However, the large quantity of *health records* hinders the flow of health information. Because each patient intervention must be documented in health records [4], we began with their statistics. In Finland with a population of approximately 5 million, about 25.5 million outpatient visits and over 7 million inpatient care days take place in public health centres every year. In Finnish public specialised care, these numbers are nearly 7.5 and 5.5 million, respectively. The numbers are increasing and supplemented by private health clinics. [5.] When studying these statistics at the broader international level, the yearly per capita number of physician consultations within the Organisation for Economic Co-operation and Development (OECD) countries is, on average, seven and the yearly hospital discharge rate per 100,000 people is over 16,000 [6].

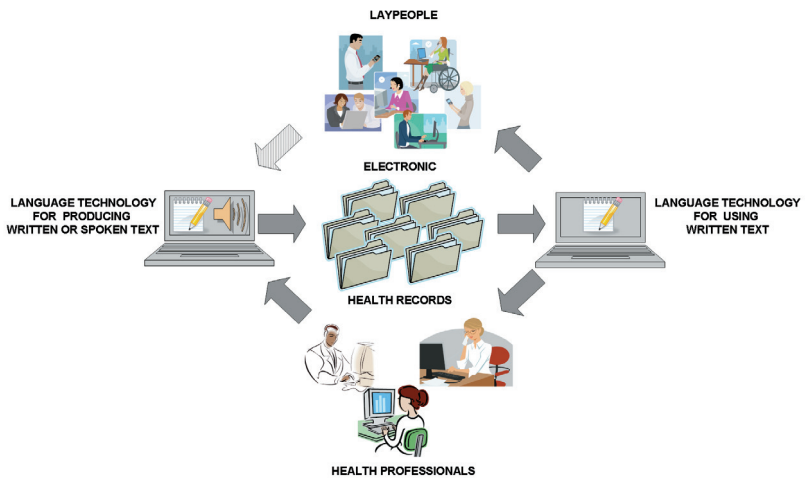

Figure 1. Language technology can support both health professionals and laypeople in producing and using textual health records [7, p. 6]. In the future, this technology can be used for information entered by laypeople themselves as indicated by the light-grey arrow.

Then, we continued with the exponentially increasing quantity of information in records per patient. With electronic health information systems, health professionals spend an average of 12–35 per cent of their working time documenting information [8, 9]. Within intensive care alone, the number of structured items that nurses type into health records has increased by 26 per cent from about 1,200 items per patient per day in 2000 to approximately 1,500 items per patient per day in 2005 [10]. These structured items are supplemented with textual information, and the amount of textual information per intensive care inpatient time can be almost 37,000 words [11]; that is, approximately 4.5 times the length of this paper. Finally, a vast amount of information is gathered automatically from various sensors and care devices. These numbers represent intensive care only, and if we combine all documentation from birth to death, from maternity clinics to pathology wards, the quantity is overwhelming and challenges the fluency of the flow of information.

At the same time, the highly specialised content of textual health records hinders the flow. About 40 per cent of health records consist of free-form text [12]. This textual part contains valuable, interpretative information on patients' status and clinical decision-making [13, 14]. However, the language of the textual records uses a highly specialised jargon and the content is very inconsistent in terms of vocabulary, grammar, structure, and topics (see, for example, [13] for data from Finnish surgical, neurological, maternity and children's wards; [15] for data from a medical-surgical ward in Thailand; [16] for data from Norwegian medicine and cardiopulmonary units; [17] for data from a US hospital; and [11] for data from Finnish and Swedish intensive care units). This complicates the use of health information: first, understanding the content is often impossible for laypeople and difficult for health professionals from another specialty or occupational group; second, only a few electronic health information systems offer tools for producing and using textual information. The lack of adequate tools for producing and using textual information becomes particularly problematic with many European languages. Consequently, creating topical overviews of textual information is difficult and time consuming and this information cannot be combined or compared with numerical or structured health information. However, such overviews, which describe trends in time, are crucial to health decision-making and create the capability to predict outcomes in healthcare and identify early indicators of possible risks.

1.2. Purpose

In this paper, we describe how computational methods for analyzing and generating natural, human language can be used to support the flow of textual information within Finnish intensive care. These methods are known as (*human*) *language technology* or *natural language processing*. Their use in supporting the flow of health information is increasingly gaining the interest of both healthcare practitioners and academic researchers. Instead of being an antithetical approach, language technology can be seen as an enabling method for developing regional and national health terminologies, standards, and other documentation tools (see, for example, [18, 19]), building multi-professional health information systems and centralised health archives (see, for example, [20–22]), as well as evaluating care quality and potential health risks (see, for example, [23, 24]). However, tailoring is needed in order to obtain language technology for the highly specialized language of health records.

More precisely, we focus on methods for building overviews of textual health records from Finnish intensive care with respect to topics specified by the user. We see this as a process whereby the user specifies the topics of interest first. Then, the related language technology would identify and highlight topically relevant text segments. The resulting overview would retain the original context and support text browsing, summarisation, and producing structured information from textual data. We address two research questions:

a) How can topical overviews be built by using machine learning methods?
b) What is the learning performance achieved in terms of statistical metrics?

1.3. Significance

This paper is timely and significant for three reasons. First, it addresses the use of intensive care data. Within intensive care, health records are typically electronic. While this constitutes a prerequisite for language technology, electronic health records are not yet used in many other hospital wards. Furthermore, the needs for a fluent flow of health information are particularly emphasised in intensive care because of critically ill patients, the overwhelming quantity of gathered information, and the highly specialised language by clinicians. Finally, language technology supporting decision making should be applicable worldwide: both the needs of end users [25] and the content of health records [11] are similar internationally.

Second, this paper proposes a novel, practise-oriented method for building clinical overviews. The method needs to produce an overview given by the topics of interest, which the user declares in advance. There are additional clinical practice-based constraints affecting the method. Almost every document contains relevant information about all topics and these topic segments are very short (on average, between four and twenty words). However, the applicability of existing methods to our problem is limited. They typically

a) require substantially longer segments (for example, approximately 200 words [26]),
b) segment the text without considering the pre-specified topics [27, 28], or
c) require a considerable amount of topic-segmented and labelled data that does not exist in relation to our problem; the model segmentation used for the training and evaluation of machine learning methods must be manually annotated.

Third, this paper addresses the application of methods for segmenting text with respect to specific topics within Finnish intensive care. The methods have been applied, for example, to English medical text from radiology and urology departments [29], English medical discharge summaries [30], and English medical dictations [31]. But, similar to many other European languages, applications for building topical overviews – or, even more generally, language technology – are still in their infancy for Finnish health records. We have studied these applications for Finnish intensive care since 2004, and in this paper, we summarize this work and address our first clinical pilot.

2. MATERIAL

The health records used in this paper consist of two datasets (Table 1). They were both collected retrospectively from Finnish intensive care units for adults. We collected the first dataset from throughout Finland in order to compare health records and their language processing nationally. With the second dataset, we studied challenges in information flow in more detail using the records of long-term patients; protracted inpatient time would most likely have increased the amount of text and, thereby, complicated the flow of information.

Table 1. Datasets from Finnish intensive care

	Dataset 1	Dataset 2
Collection year	2001	2006
Inclusion criteria	3 patients from every intensive care unit	Intensive care inpatient period ≥ 5 days between 1 Jan 2005 and 1 Aug 2006
No. of hospitals	15	1
No. of patients	43 (2–3 per hospital)	516 (18.5 per cent of all patients)
Documents	For every patient: - Admission document - Daily notes for one 24 h period	Admission documents: - 348 documents - 87,000 words Daily notes: - documents for 516 patients - 18,400 shift-related documents - 1.1 million words Discharge documents: - 514 documents - 206,000 words
No. of words	5,100	1.393 million
Further details	[7, pp. 16, 17, 52, 53]	[7, pp. 16–34, 66]

For both datasets, we received the proper permits from ethical committees and hospital authorities. Patient confidentiality was assured by using anonymous document identifiers and maintaining patient privacy. When tailoring commercial language technology to the research domain, we manually de-identified the material provided to commercial parties.

Both datasets included text documents written by intensive-care nurses. Every patient's *nursing records* were grouped, according to the time of writing, into *admission documents*, *daily notes*, and *discharge documents* (Figure 2). We chose nursing records because they cover the entire intensive care inpatient period and because other professionals write substantially less in Finnish intensive care. Consequently, we considered nursing records as the most challenging records in terms of streamlining the fluent flow of health information.

In accordance with our framework of supporting communication and the decision making of health professionals and laypeople (see Section 1.1), we intend to extend the study to cover health text written by different professional groups and laypeople, to healthcare domains other than intensive care, and to languages other than Finnish. We have already begun this extension work by comparing nursing records from Finnish and Swedish intensive care units [11].

ADMISSION DOCUMENT (compact patient description at the beginning of the intensive care inpatient period)

PREVIOUS/OTHER DISEASES: BP disease, Chr. FA

ANAMNESIS: 18.3 bad chest pain starts. Strong ST changes in Loimaa. Angiography and discovery of a tight main stenosis. LAD good, substantial changes in RCA and LCX. Edema and intubation in the morning of 19.3 in TUH. *NAME* infusion started. ICU ad. for an emergency operation. O room: Low pressures before perfusion. A balloon pump will be placed, *NAME* goes in. Pulmonal pressures high -> *NAMES*. At the end of perfusion had to start *NAME* infusion too. Got 4 x fresh frozen plasma and thrombocytes.

Dg: MCC Oprtn: Reconstr. coron. cordis No. IV

¤ Ao -> *NAME* 140ml/h

¤ Ao -> *NAME* 154 ml/h

DAILY NOTES (shift-related documents from intensive care inpatient period, 2 shift-related documents below)

Long night s. BP tried st. easily rise even to very high values towards midnight, towards morning BP level went down and became stabile. p slightly tachycardic. Profuse diuresis. Towards midnight fillingpressures occ. highish, towrads morning went down subst. Suff CI (c. level 2), rose ad 2.5 at morning. With 40 % vm .oxidation on the low side, ventilates well. Put a 50 % mask with outcome good ox. Br.exercises with benett go well. The wound of the r. leg dripped pl. of tissue fluid st., dressings changed to the bottom once, Ext. tired, opens eyes occ. and tries to answer to the posed questions but lacks the strength often.

long shift

hemod: BP mainly high and pulse tachy still. Pulse occasionally sinus rhythm. Before noon got 2 ampoules of *NAME* every hour; this made pulse only slightly more stabile. In the afternoon, BP down very strongly after *NAME*.Pulmonal catheter rmved

breathing: increased pleural fluid in the x-ray; "drying" up the patient continues heavily. Oxidation improved during the day diuresis been very profuse, *NAME* cont with the previous dose because of the lung situation

consciousness: in the morning been very tired again, during the forenoon perked up a bit. But still not up to talk lots. During the afternoon has started to faint and finger tubes. Got *NAME* 5 mg iv. which decreased the pressure substly but made absolutely shattered. Extr. shattered also after *NAME* excretion of the right leg has smwht decreased, edema throughout too per os taken some drinks and a little bit of gruel in the afternoon

DISCHARGE DOCUMENT (compact patient description at the end of the intensive care inpatient period)

REASON FOR THE ICU ADMISSION AND ANAMNESIS *Verbatim copy of the admission document, except O replaced with operation.*

BREATHING: Ox problems at the beginning, when the situ improved started to wean

22.3. Extubated

23.3. before noon Thick, yellow mucus from tube, sufficient extubation with the 40/50% ventimask.

24.3 with 28% VM oxygen 9.5 and CO2 5.6. 23.3 in the chest x-ray increased pleural fluid and mild incompensation.Continued hevy minussing (*NAME* 5mg x 6 iv.) The amount of pleural fluid dropped. *NAME* stopped at11, will be given when needed according to the response.

HAEMODYNAMICS: Pulse tachycardic,extrasystoles, flimmer. Admission BP low. *NAMES* infusions of a large dosage. Filling. However, CI quite low. IABP 1:1. 20.3. started *NAME*, when the rhythm did not convert with shock. 21.3. *NAME* stopped As a new finding, left branch block, which improved 20.3 C.I 1.5 --> started *NAME* infusion. Currently BP even too high. Pulse: FA, tachycardia.

CONSCIOUSNESS AND MOOD: ICU started *NAME* infusion, got *NAME* boluses too. *NAME* stopped 23.3. after extubation, afterwards still very tired, but trying to kooperate, however. 25.3. tired and speeking is difficult, answers with single words. Slightly towards perking up.

NUTRITION: Small portions of liquid bo. taken.

EXAMINATIONS:

¤ 18.3 esophagus -US performed by H. Suominen: A clear bag of fluid around the heart , good contraction.

¤ 20.3 esophagus- US / K. Haverinen: Septum is faint, contracts mod. Mitral- and aortic valve ok. No explanation for L block been found.

INFECTION SITUATION: 23.3 CRP 32

EXCRETION: Profuse diuresis, because *NAME* 10 mg iv every 4h.26.3 Diuresis at 6 -->: 12 ->1090 ml

Motion: -

Drains: After o profuse drain-bleeding, Hb low, needed a lot of red blood cs.In the esophagus-US observed a bag of fluid around the heart and impending tamponation. Obsed the situation until next 8 a.m., when the cndsn was resternomy, where found 2 locations of bleeding, which fixed. O bleeding 750 ml

FLOW: AO-LAD 75 ml/min AO-LOM-LPL 120 ml/min AO-RCA 180 ml/min After the resternotomy the bleeding went down and changed to serous. Drains (x 3; in front of the heart + both pleuras) removed 22.3

SKIN CONDITION AND CARE: R. leg wound bled a lot. Bandage changed at least once a day. Stitched up from the point of bleeding on 15.3. Sternum wound tidy. Left buttock has a decubitus ulcer; 15.3. started caring with *NAMES*. Skin sensitive all over; nicks caused by tapes etc.

PAIN MANAGEMENT: *NAME*, which makes very tired.

SPECIAL CARE: 19.-23.3. IABP

RELATIVES: Two sons been in contact, also the male friend visited. Son has called and knows about the discharge to Loimaa.

BELONGINGS: Clothing bag collected from the w. Two children's drawings and a card -> put into the clothing bag.

OTHER INSTRUCTIONS Blood s. vary -> *NAME* infusion. Insulin interrupted for the transit. Potassium high and changed PL-K to Na 0.3 at 12.

Figure 2. Anonymous English translation of health records from a Finnish intensive care unit. The style, including typographical errors, was preserved. In order to be free of commercialism, pharmaceutical trademarks have been replaced with *NAME*.

3. METHODS

We begin by considering topical search. First, we simplify the learning task to the *identification of topically relevant segments* from a collection of text segments. The search topics are assumed to be independent of each other in order to allow a segment to be relevant for multiple topics. Second, we supplement the identification task with a *topical relevance evaluation*. For every topic, we use a continuous scale ranging from irrelevance to maximal relevance. Third, we consider the entire task of *segmenting text with respect to specific topics*, both the identification of boundaries where the topic changes and assignment of the respective topic labels with the assumption of assigning each text segment to the most relevant topic. Finally, in order to improve the performance of the topical search system and to make the output text easy to understand, we develop language technology for the linguistic analysis of health records.

3.1. Topical search

Machine learning methods were applied to the topical search in this work. We began by using *supervised* machine learning techniques, where we first annotated data in order to define the desired outcome and then let the method learn a function that connects text to this outcome. We tailored these methods then into an *unsupervised* direction in order to meet the user's need for an *ad hoc* information search without fixing the set of possible search topics in advance. We described the methods in more detail in [32], [33], and [34] for sections 3.1.1, 3.1.2, and 3.1.3, respectively.

3.1.1. Multi-label classification

Our task was to identify topically relevant segments from a set of text segments known as a *binary text classification problem*. The input was a text segment and the search system classified it as topically relevant or irrelevant. Because we considered each search topic independently, we had a *multi-label classification problem*.

We used the *regularized least-squares algorithm* [35] as our learning method. The algorithm learned the function for connecting text segments with the labels of topically relevant and topically irrelevant by minimizing the distance between the desired labels and the function values as well as by controlling the complexity or smoothness of the function simultaneously.

With our first dataset, we evaluated the learning quality using the topics of *breathing*, *blood circulation*, and *pain*. We chose these topics in collaboration with a senior researcher in clinical nursing science for two reasons. First, their monitoring is crucial in intensive care. Nurses evaluate patients' care needs regularly, and use textual health records in this task to support their decision making. In Finland, the evaluations are based on a model of intensive care nursing in which our three topics play a central role [36]. Second, the topics are different from a machine-learning-ability viewpoint due to textual differences in breathing, blood circulation, and pain documentation. A content expert first divided the dataset into text segments that described only one matter at a time. Then this expert and two other experts annotated them with respect to the three topics. After initial annotation training, the experts worked independently of each other. The first two experts were registered nurses with lengthy clinical experience in

intensive care nursing and the third was the aforementioned senior researcher in clinical nursing science.

In summary, our task included the following:

Input

a) Topics of interest considered independent of each other.
b) A set of text segments manually classified with respect to these topics (in other words, *expert annotation*).

Training and testing

a) Divide the annotated data into training and test sets.
b) Use training data to specify the function in the regularized least-squares algorithm.
c) Use the trained system to process text segments in the test set.
d) Evaluate the learning performance by comparing the output topics with the annotation.

3.1.2. Multi-label regression

Our task was to evaluate the topical relevance of each text segment using a continuous scale, known as a *text regression problem*. Again, we considered every topic independently (in other words, a *multi-label regression problem*) and used the regularized least-squares algorithm but now in a regression model. The task summary was identical to that of the multi-label classification (section 3.1.1).

The text segments were taken from our first dataset. The topics were *breathing*, *blood circulation*, and *pain*. Learning and evaluation were based on manually annotated data, in which we assigned a regression score of zero to irrelevant text segments and scores of one, two and three for the increased relevance to the topic. As an example, the text segment '*Kipulääke auttaa korkeaan pulssiin.*' ['*A pain killer is the cure for a high pulse.*'] received a pain score of three because it provides important information on the patient's pain status and efficiency of pain management. The segment '*kääntöä vastustaa*' ['*resists turning over*'] received a pain score of one because it indicates possible pain during or after an intervention. Scores were determined based on the annotation for multi-label classification (section 3.1.1); the score was equal to the number of content experts who annotated the segment as relevant to a given topic.

3.1.3. Multi-class classification

Our task was to segment text with respect to specific topics. This is known as a *multi-class classification problem*. For each word of input text, the language technology component assigned one topic from a set of possible topics.

We employed the *Hidden Markov model* [37] as our learning method. First, we considered it in a standard supervised method; an optimal topic sequence for an input text sequence was found through learning from manually annotated data. More specifically, the model learnt the probabilities for assigning a given topic to a certain word (in other words, *emission probabilities*) and the probabilities for observing a given topic when the previous topic is known in advance (in other words, *transition probabilities*). Then, we tailored this method into an unsupervised direction by

specifying emission probabilities through comparing textual contexts in which the topic keyword and the word in question usually occur. For transition probabilities, we used a simple parameterisation by controlling *self-transitions*, that is, assignments of the same topic to two consecutive words with a parameter δ, which takes values between zero and one. We distributed the remaining probability $1 - \delta$ of switching the topic uniformly between other topics.

For learning and evaluation, we used the second dataset. Two content experts collaborated and created one manual annotation using the most common topics from our dataset – *breathing*, *hemodynamics*, *consciousness*, *relatives*, and *diuresis* – as well as the topic of *other* that is used for words irrelevant to each of these five topics. We chose the topics in collaboration with clinical nursing science researchers. These topics were chosen because they play a central role in the model for monitoring Finnish intensive care patients [36] and because their use as documentation topics was well established in the intensive care unit where the data were collected. The annotators discussed topic segmentation and labelling decisions until they reached a consensus.

In summary, our task included the following:

Input

a) Topics of interest (and the topic of *other*).
b) Annotated text, where each word is associated with one of the topics.
c) For the unsupervised method, using unannotated text is also possible.

Training and testing

a) Divide the annotated data into training and test sets.
b) Use training data to specify the probabilities of the supervised hidden Markov model.
c) Use the trained system to process text in the test set.
d) Use the unsupervised system to process text in the test set.
e) Evaluate the performance of supervised and unsupervised methods by comparing the output segmentation with the annotated segmentation.

3.2. Linguistic analysis: parsing and tailoring

As a supporting step for machine learning, we considered language technology for linguistic analysis. More precisely, we developed new methods for parsing, and tailored commercial state-of-the-art software for language related to health records by using daily notes from our second dataset in [38–40] (we refer to these original papers for detailed descriptions of methods and their evaluation). Our parsing methods included both rule-based and statistical approaches. The tailoring work extended the vocabulary of Lingsoft's Finnish morphological analyser [41]. This was performed in close collaboration with Lingsoft by combining their expertise in language technology for Finnish and our expertise in developing methods for text from biomedical papers and electronic health records.

4. RESULTS AND DISCUSSION

4.1. Topical search

4.1.1. Multi-label classification

We considered the multi-label classification problem using three topics and 1,367 text segments, as well as three content expert annotations in [32]. The average segment length was 3.7 words. On average, the content experts annotated twenty per cent, fifteen per cent, and six per cent of the segments as relevant to the topics of *breathing*, *blood circulation*, and *pain*, respectively. Their agreement on topic classification was substantial, as shown in Table 2.

The classification performance was promising (Table 2). We obtained the best results from the topics of *breathing* and *blood circulation*. The topic of *pain* had a considerably smaller amount of positive training instances (approximately 19 per cent of the training instances compared to approximately 20 per cent for *breathing* and *blood circulation*).

Table 2. Results of multi-label classification, including the inter-annotator agreement with the data from 1,363 text segments, and the classification performance with the training data from 708 segments and the test data from 655 segments [7, pp. 55, 56]

Inter-annotator agreement Cohen's κ [42] (95% confidence interval)			
	Breathing	Blood circulation	Pain
Content experts 1 and 2	0.73 (0.68–0.78)	0.89 (0.85–0.92)	0.88 (0.82–0.94)
Content experts 1 and 3	0.67 (0.62–0.72)	0.81 (0.77–0.86)	0.79 (0.73–0.86)
Content experts 2 and 3	0.85 (0.82–0.89)	0.87 (0.83–0.90)	0.76 (0.69–0.83)
Classification performance AUC [43] (95% confidence interval)			
	Content expert 1	Content expert 2	Content expert 3
Breathing			
Classifier trained for content expert 1	0.86 (0.82–0.90)	0.74 (0.69–0.79)	0.72 (0.68–0.77)
Classifier trained for content expert 2	0.83 (0.79–0.88)	0.88 (0.85–0.91)	0.86 (0.83–0.89)
Classifier trained for content expert 3	0.84 (0.80–0.88)	0.88 (0.84–0.91)	0.87 (0.84–0.91)
Blood circulation			
Classifier trained for content expert 1	0.89 (0.84–0.93)	0.93 (0.89–0.97)	0.91 (0.87–0.95)
Classifier trained for content expert 2	0.88 (0.83–0.93)	0.93 (0.90–0.97)	0.91 (0.86–0.95)
Classifier trained for content expert 3	0.89 (0.84–0.93)	0.93 (0.90–0.97)	0.91 (0.86–0.95)
Pain			
Classifier trained for content expert 1	0.71 (0.61–0.80)	0.81 (0.72–0.90)	0.72 (0.63–0.81)
Classifier trained for content expert 2	0.71 (0.61–0.80)	0.81 (0.73–0.89)	0.71 (0.62–0.80)
Classifier trained for content expert 3	0.67 (0.56–0.78)	0.77 (0.66–0.87)	0.71 (0.61–0.80)

4.1.2. Multi-label regression

We supplemented the multi-label classification problem with the topical relevance evaluation in [33]. We trained our regression system with the combined expert annotation. Again, the results were promising, with a moderate performance for the topic of *pain* (Table 3). The system recognized the most relevant segments well, but more implicit indicators were difficult to identify automatically, again due to having fewer training instances with the respective characteristics.

Table 3. Results of multi-label regression, including the number of text segments with different scores in training and testing, as well as the regression performance, with the training data from 708 text segments and the test data from 655 segments [7, pp. 57, 59]. Note that the performance measure values range in the interval of [-1, 1] instead of [0, 1] used in Table 2.

Number of segments Training/Testing			
	Breathing	Blood circulation	Pain
Annotated score 0	572/462	564/566	646/609
Annotated score 1	28/31	26/15	14/14
Annotated score 2	35/60	23/9	7/10
Annotated score 3	73/102	95/65	41/22
Regression performance Kendall's *τ-b* [44, p. 40] (95% confidence interval)			
	Breathing	Blood circulation	Pain
Kendall's *τ-b* (95% confidence interval)	0.62 (0.56–0.68)	0.69 (0.61–0.76)	0.44 (0.30–0.59)

To illustrate the potential of a relevance evaluation for building overviews, we considered the need to quickly develop an overview for issues related to blood circulation ([7, pp. 58, 59], Figure 3). The user could select a sensitivity level (for example, fifty or one-hundred segments). We used the regression system to return the corresponding number of text segments in an ascending order of relevance. We performed the experiments using the topic of *blood circulation* and our test set.

With a sensitivity level of fifty text segments, a great majority (that is 45 segments) of automatically returned text segments had an annotated score of three. The number of returned segments with a blood circulation score of one and two in the manual annotation were one and two, respectively. Only two returned segments ('*tilanne stabiili*' ['*situation stable*'] and '*nostettu infuusiota*' ['*lift in infusion*']) were considered irrelevant in the manual annotation. In comparison, the number of segments with an annotated blood circulation score of zero, one, two, and three in the test set were 566, fifteen, nine, and 65, respectively (Table 3).

With a sensitivity level of one-hundred text segments, 55 segments with a score of three in the manual annotation were returned. The number of returned segments with a score of one and two in the manual annotation were one and five, respectively. However, the number of segments that were considered irrelevant in the manual annotation increased to 39. If all segments with blood circulation scores one, two, and three in the manual annotation had been returned, the number of topically irrelevant segments should have been eleven for the system output of one-hundred segments.

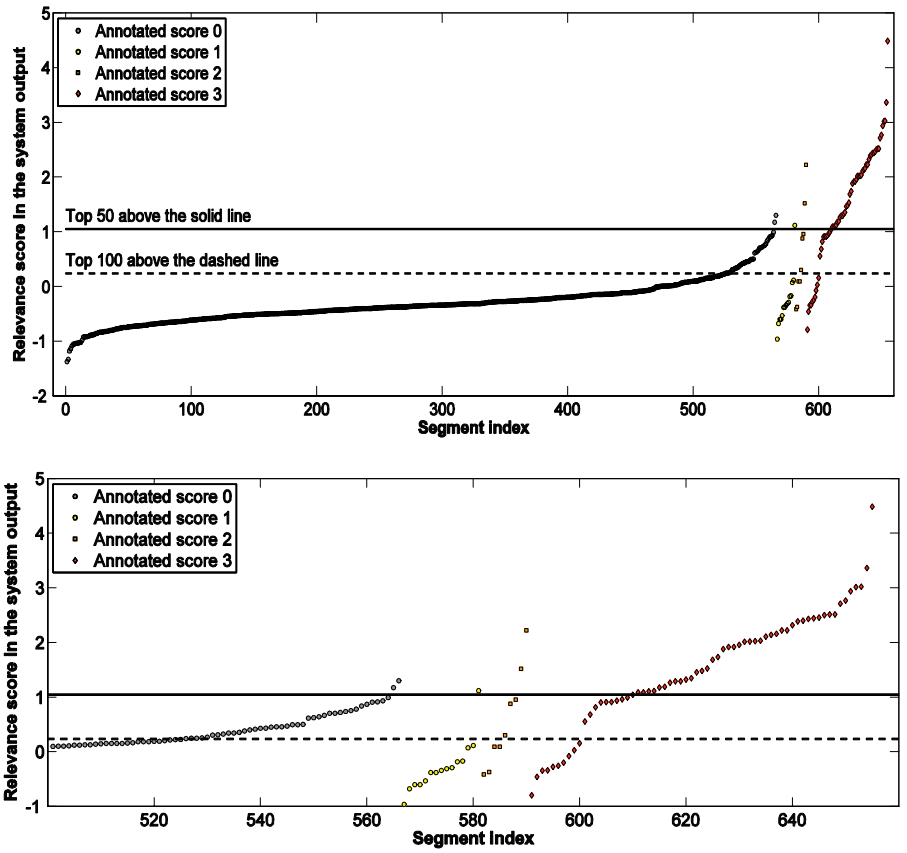

Figure 3. Regression using the topic of *blood circulation* and a test set of 655 text segments. The training data consisted of 708 segments. The segment indices refer to individual text segments in our test set, organised separately for each of the four expert annotation score values based on an increasing order of relevance with respect to the system output. The upper figure includes all 655 text segments, and the portion for segment index > 500 is magnified in the lower figure.

4.1.3. Multi-class classification

We addressed the entire task of topic segmentation and labelling (Figure 4) in [34]. At least four of the five search topics (that is to say, *breathing*, *hemodynamics*, *consciousness*, *relatives*, *diuresis*) were discussed in over sixty per cent of the 405 shift-related documents in the annotated data. The proportion of documents that contained none of them was about eight per cent. The average segment length was eighteen words, but it varied substantially with the search topics (for example, segments related to the topic of *hemodynamics* were typically long, while those related to the topic of *diuresis* included only a couple of words). Taking this segment-level characteristic into account was crucial in method development.

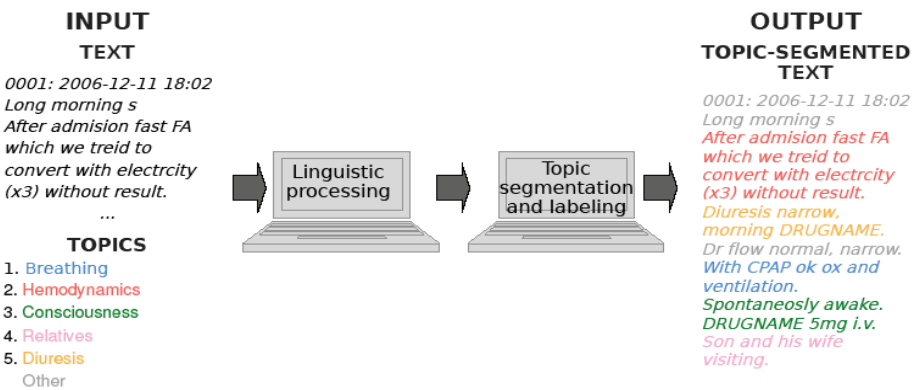

Figure 4. The process of creating topical overviews. The input consists of free-form text and topics of interest. After linguistic processing, followed by topic segmentation and labelling, the system outputs the topic-segmented text. Similarly to Figure 2, this figure illustrates the style, including typographical errors, of health record text from a Finnish intensive care unit.

We performed a performance comparison of our topic segmentation and labelling methods (Figure 5). The performance of the supervised method was good. As expected, it was better than that of our new, unsupervised method when we used all available data for the training. If we used all data for the training, the classification accuracy of the supervised method was 0.83 on the test set. The accuracy of the new, unsupervised method on the test set was 0.75 (see Figure 5), substantially better than that of an unsupervised baseline heuristic, 0.67[+]. The supervised method outperformed the unsupervised one if at least 3,600 words were used for training. This translates to annotating approximately fifty shift-related documents with an average length of eighty words. The numbers for the unsupervised baseline heuristic were 2,000 words or 25 documents. The accuracy of the supervised method increased substantially if we used approximately 8,000 words (about one-hundred documents) for the training.

[+]This heuristic inherently resembled the structure of our data. It identified topic keywords (for example, *breathing*) from the text and assigned each word to the topic indicated by the previously seen keyword. The topic was initialized at the beginning of a document as *other*.

Linguistic processing contributed to the performances of all three methods (Figure 5). Its importance was more significant if we used fewer data for the training. For every word recognized by the tailored analyser, we used the first output lemma. For example, for the Finnish word *haavan* [*of the wound*], the output included the lemmas of *haapa* [*aspen*] and *haava* [*wound*]. For words not recognized by the tailored analyser, we preserved the original spelling. Our results indicate that this processing reduced the problems related to the sparseness of highly inflectional Finnish, and our systematically similar selection of the lemma mapping did not harm an automated understanding of the text's meaning.

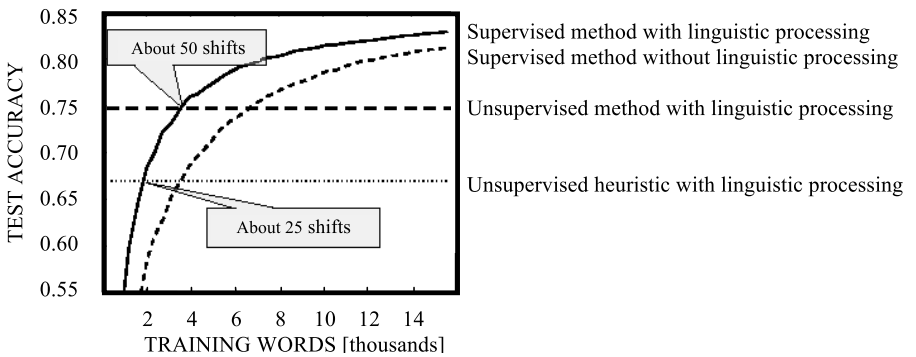

Figure 5. Comparison of topic segmentation and labelling methods using the test set of 204 shift-related documents and increasing the training set up to 16,000 words

4.2. Linguistic analysis

4.2.1. Parsing

We previously developed rule-based and statistical parsers for processing textual health records [38, 39]. This resulted in not only new methods but also important information on comprehensibility and the content of the records. Without parsers (or substantially larger datasets), the knowledge stated in textual health records was not readily available for other language technology components. For example, negations and speculations were common, and parsing was needed to identify whether a certain medicine was given or not or whether a certain diagnosis was confirmed or speculated.

By comparing our rule-based and statistical parsers, we concluded, that even with our relatively small corpus, our statistical parser achieved results comparable to previous studies with considerably larger datasets. This conclusion proved true for a number of languages.

We also created a set of rules for transforming the constituency scheme of our rule-based parser to fit the dependency scheme of our manually annotated dataset for a syntax analysis. This improved the applicability of our rule-based parser by deepening its analysis results with inferred grammatical roles.

4.2.2. Tailoring

We tailored the vocabulary of the analyser by extending approximately 3,500 clinical terms. The standard vocabulary covered approximately 85 per cent (900,000 words) of daily notes in our second dataset. The extension included all words not recognized by the analyser which occurred at least fifty times in the daily notes of our second dataset. The five most common words were *Diureesi* ([*Diuresis*], with about 6,300 occurrences), *Hemodynamiikka* ([*Hemodynamics*], with about 5,200 occurrences), *RR* (ambiguous abbreviation, with about 3,900 occurrences), *HEMODYNAMIIKKA* ([*HEMODYNAMICS*], with about 3,600 occurrences), and *SR* (abbreviation for *sinusrytmi* [*sinus rhythm*], with about 3,300 occurrences). The list continued with professional jargon, the names of medicines, interventions and laboratory tests, abbreviations, and frequent misspellings.

The extension substantially improved the applicability of the analyser to the health domain. With our daily notes, it resulted in a relative gain of 42 per cent in texts covered by the analyser [38]. The success led to piloting of the present language technology components in an authentic healthcare environment in the autumn of 2008. The pilot project included basic components for producing and using textual health records: linguistic and stylistic proof-reading, domain-terminology building, and aids in understanding (that is to say, links to dictionaries and terminologies). The results confirmed that health professionals perceive the tools as useful. [7, pp. 101–103.] This led to the release of a commercial language technology for Finnish health records in 2009 [40].

5. CONCLUSION

In conclusion, the results from building overviews were promising for real-life applications. In terms of the problem of identifying text segments relevant to topics of *breathing*, *blood circulation*, and *pain*, our supervised classifier was able to learn the task. The classifier also had a relatively good generalisation capability for multiple content experts' opinions.

Our regression system was able to distinguish the segments most relevant to the topic of *blood circulation*. Although supervised learning was possible with a relatively small number of topically relevant segments, the system performance was limited with certain topics (for example, *pain*). Consequently, further development is needed to meet the quality requirements of clinical decision making in intensive care.

Regarding simultaneous topic segmentation and labelling with respect to topics of *breathing*, *hemodynamics*, *consciousness*, *relatives*, and *diuresis*, the results from our supervised system were encouraging in terms of statistical metrics. Furthermore, this type of structuring was shown to increase the information search speed of health professionals [45]. We also addressed building overviews with *ad hoc* search topics by developing an unsupervised method and evaluating its performance against the annotation work needed to train the supervised system.

6. FUTURE WORK

In the future, tools for building overviews have the potential to enhance the quality and efficiency of care by improving access to health information. The tools will allow health

professionals more time for direct care. This will also have positive impacts on the efficiency and profitability of health services. Linguistic analyses will also support clinicians' legal obligations to produce high-quality health records and are in line with the richness of unit-specific documentation practices. Moreover, the approach of building topic-specific overviews enables us to take the target audience into account; for example, patients, various professional groups such as physicians, nurses, and physiotherapists within the intensive care unit, and the doctors at the hospital ward to which the patient is discharged all have different needs. The same technology can be applied to the records of patient populations. This makes it possible to analyze large numbers of textual health records systematically which enables health researchers to use this cumulative, knowledge-rich resource to enrich clinical evidence base. More discussion on the potential of language technology to support information flow in intensive care can be found in [46] and [7, pp. 15–42 and pp. 97–106].

Our methods for building overviews are likely to be applicable to intensive care units worldwide since both the needs of end users [25] and the contents of health records [11] are similar in other countries. However, in addition to linguistic tailoring, we need more cross-language method comparisons such as that for English and Portuguese health records [47]. We also need more real-life case studies that demonstrate the usability and usefulness of the methods and generalizability of the search topics. Examples of studying language technology in clinical practise can be found in [48–50]. The increased availability of de-identified health records and open-source methods for processing textual health information support efforts related to these needs [51, 52].

Finally, in order to establish the methods in clinical practise, their commercialisation as well as integration within existing healthcare and communication processes should be addressed. This holistic approach includes connecting language technology components to their developers. For example, during the language technology development described in this paper, we have established a living link between healthcare providers, health terminology developers, software houses in hospital information systems and language technology, as well as research groups in health informatics. The outcome of the collaboration included

a) the release of commercial software for processing Finnish health records,
b) a platform for the easy integration, tailoring, and extension of language technology to existing health information systems, and
c) an establishment of Finnish, Nordic and Baltic, as well European clusters [53, 54].

Each of these clusters aims to support the production and use of health records by developing language technology solutions, performing domain comparisons and delivering and applying regional results to other regions.

ACKNOWLEDGEMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. We also received funding from the Finnish Funding Agency for Technology and Innovation, Tekes (grants 40435/05, 40020/07) and Academy of Finland (decision 136653). We express our gratitude to RN, MNSc Heljä Lundgrén-Laine (University of Turku, Department of Nursing Science and

Hospital District of Soutwest Finland); Dr Leif Hanlen (NICTA, Canberra Research Laboratory and Australian National University, College of Engineering and Computer Science); all the members of the Louhi project, IKITIK consortium, and HEXAnord network; as well as to the anonymous reviewers.

REFERENCES

[1] Glaser, S.R., Zamanou, S. and Hacker, K., Measuring and Interpreting Organizational Culture, *Manag Comm Q*, 1987, 1(2), 173–198.

[2] US Department of Health & Human Services, Health Insurance Portability and Accountability Act of 1996, HIPAA 1996, http://www.cms.hhs.gov/HIPAAGenInfo/Downloads/HIPAALaw.pdf [accessed 2009, November, 18].

[3] Mills, M.E., Linkage of Patient Records to Support Continuity of Care: Issues and Future Directions, *Stud Health Techn Inform*, 2006, 122, 320–324.

[4] Statutes of Finland, Decree 99/2001 of the Ministry of Social and Health, Finland, http://www.finlex.fi [accessed 2009, November, 18].

[5] Stakes, The Statistical Yearbook on Social Welfare and Health Care 2008, Yliopistopaino, Helsinki, Finland, 2008.

[6] OECD.Stats Extracts, OECD Health Data 2009, http://stats.oecd.org/Index.aspx?DatasetCode=HEALTH [accessed 2009, November 18].

[7] Suominen, H., Machine Learning and Clinical Text: Supporting Health Information Flow, *TUCS Dissertations*, 2009, 125.

[8] Hakes, B. and Whittington, J., Assessing the Impact of an Electronic Medical Record on Nurse Documentation Time, *J Crit Care*, 2008, 26(4), 234–241.

[9] Banner, L. and Olney, C.M., Automated Clinical Documentation: Does It Allow Nurses More Time for Patient Care? *J Crit Care*, 2009, 27(2), 75–81.

[10] Manor-Shulman, O., Beyene, J., Frndova, H. and Parshuram, C., Quantifying the Volume of Documented Clinical Information in Critical Illness, *J Crit Care*, 2008, 23(2), 245–250.

[11] Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Vidas, D., Hassel, M., Kokkinakis, D., Lundgrén-Laine, H., Nilsson, G., Nytrø, Ø., Salanterä, S., Skeppstedt, M., Suominen, H. and Velupillai, S., Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives, *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi 2010)*, Los Angeles, USA, 2010, 56–60.

[12] Dalianis, H., Hassel, M. and Velupillai, S., The Stockholm EPR Corpus – Characteristics and Some Initial Findings, *Proceedings of The 14th International Symposium for Health Information Management Research, ISHIMIR-09*, Kalmar, Sweden, 2009.

[13] Kärkkäinen, O. and Eriksson, K., Evaluation of Patient Records as Part of Developing a Nursing Care Classification, *J Clin Nurs*, 2003, 12(2), 198–205.

[14] Suominen, H., Lundgrén-Laine, H., Salanterä, S., Karsten, H. and Salakoski, T., Information Flow in Intensive Care Narratives, *Proceedings IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBM 2009*, Washington DC, USA, 2009, 325–330.

[15] Cheevakasemsook, A., Chapman, Y., Francis, K., and Davies, C., The Study of Nursing Documentation Complexities, *Int J Nurs Pract*, 2006, 12(6), 366–374.

[16] Hellesø, R., Information Handling in the Nursing Discharge Note, *J Clin Nurs*, 2006, 15(1), 11–21.

[17] Hyun, S. and Bakken, S., Towards the Creation of an Ontology for Nursing Document Sections: Mapping Section Headings to the LOINC Semantic Model, *AMIA Annu Symp Proc*, 2006, 364–368.

[18] Zhou, L., Tao, Y., Cimino, J.J., Chen, E.S., Liu, H., Lussier, Y.A., Hripcsak, G. and Friedman, C. Terminology Model Discovery Using Natural Language Processing and Visualization Techniques, *J Biomed Inform*, 2006, 39(6), 626–636.

[19] Deléger, L., Merkel, M. and Zweigenbaum, P., Translating Medical Terminologies through Word Alignment in Parallel Text Corpora, *J Biomed Inform*, 2009, 42(4), 692–701.

[20] De Clercq, E., Problem-oriented Patient Record Model as a Conceptual Foundation for a Multi-professional Electronic Patient Record, *Int J Med Inform*, 2008, 77(9), 565–575.

[21] Silvester, B.V. and Carr, J.S. A Shared Electronic Health Record: Lessons from the Coalface, *Med J Aust*, 2009, 190(11), S113–S116.

[22] Virtanen, T. The Finnish National eHealth Archive and the New Research Possibilities, *Stud Health Techn Inform*, 2009, 146, 688–691.

[23] Baldwin, K.B., Evaluating Healthcare Quality Using Natural Language Processing, *Healthc Qual*, 2008, 30(4), 24–29.

[24] Hripcsak, C., Soulakis, N.D., Morrison, F.P., Lai, A.M., Friedman, C., Calman, N.S. and Mostashari, F., Syndromic Surveillance Using Ambulatory Electronic Health Records, *J Am Med Inform Assoc*, 2009, 16(3), 354–361.

[25] Lauri, S. and Salanterä, S., Developing an Instrument to Measure and Describe Clinical Decision Making in Different Nursing Fields, *J Prof Nurs*, 2002, 18(2), 93–100.

[26] Hearst, M.A., TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Comp Ling*, 1997, 23(1), 33–64.

[27] Ponte, J.M. and Croft, W.B., Text Segmentation by Topic, *LNCS*, 1997, 134, 113–125.

[28] Chang, T.-H. and Lee, C.-H., Topic Segmentation for Short Texts, *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, PACLIC*, 2003, Singapore, Singapore, 159–165.

[29] Cho, P.S., Taira, R.K. and Kangarloo, H., Automatic Section Segmentation of Medical Reports, *AMIA Annu Symp Proc*, 2003, 155–159.

[30] Bramsen, P., Deshpande, P., Lee, Y. K. and Barzilay, R., Finding Temporal Order in Discharge Summaries, *AMIA Annu Symp Proc*, 2006, 81–85.

[31] Jancsary, J. and Matiasek, J., Revealing the Structure of Medical Dictations with Conditional Random Fields, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2008, Honolulu, Hawaii, 1–10.

[32] Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S. and Salakoski, T., Towards Automated Classification of Intensive Care Nursing Narratives, *Int J Med Inform*, 2007, 76(S3), S362–S368.

[33] Suominen, H., Pahikkala, T., Hiissa, M., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S. and Salakoski, T., Relevance Ranking of Intensive Care Nursing Narratives, *LNCS*, 2006 4251(1), 720–727.

[34] Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T., Combining Hidden Markov Models and Latent Semantic Analysis for Topic Segmentation and Labeling: Method and Clinical Application, *Int J Med Inform*, 2009, 78(12), e1–e6.

[35] Poggio, T. and Smale, S., The Mathematics of Learning: Dealing with Data, *Notices of the American Mathematical Society (AMS)*, 2003, 50(5), 537–544.

[36] Fagerström, L., Rainio, A. K., Rauhala, A., and Nojonen, K., Validation of a New Method for Patient Classification, the Oulu Patient Classification, *J Adv Nurs*, 31(2), 481–490.

[37] Rabiner, L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proc IEEE*, 1989, 77(2),257–286.

[38] Laippala, V., Ginter, F., Pyysalo, S. and Salakoski, T., Towards Automated Processing of Clinical Finnish: Sublanguage Analysis and a Rule-based Parser, *Int J Med Inform*, 2009, 78(12), e7–e12.

[39] Haverinen, K., Ginter, F., Laippala, V. and Salakoski, T., Parsing Clinical Finnish: Experiments with Rule-Based and Statistical Dependency Parsers, *NEALT Proceedings Series*, 2009, 8, 65–72.

[40] Lingsoft, Lingsoft julkisti kielentarkistimen terveydenhuollon kielelle [Lingsoft Released a Proof-reading Program for Health Care Jargon], press release, 2009 April 28, http://www.lingsoft.fi/?doc_id=438&lang=fi [accessed August 1, 2009].

[41] Koskenniemi, K., Two-level Model for Morphological Analysis, *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 1983, 683–685.

[42] Cohen, J., A Coefficient of Agreement for Nominal Scales, *Educ Psychol Meas*, 1960, 20(3), 37–46.

[43] Hanley, J.A. and McNeil, B.J., The MEaning and Use of The Area Under a Receiver Operating Characteristics (ROC) Curve, *Radiology*, 1982, 143(1), 29–36.

[44] Kendall, M. and Gibbons, J.D., *Rank Correlation Methods, 5th edition*, Edward Arnold, London, UK, 1990.

[45] Tange, H.J., Schouten, H.J., Kester, A.D.M and Hasman, A., The Granularity of Medical Narratives and Its Effect on the Speed and Completeness of Information Retrieval, *J Am Med Inform Assoc*, 1998, 5(6), 571–582.

[46] Raja, U. Mitchell, T., Day, T. and Hardin, J.M., Text Mining in Healthcare. Applications and Opportunities, *J Healthc Inf Manag*, 2008, 22(3), 52–56.

[47] Castilla, A.C., Furuie, S.S. and Mendonça, E.A., Multilingual Information Retrieval in Thoracic Radiology: Feasibility Study, *Stud Health Techn Inform*, 2007, 129, 387–391.

[48] Mendonça, E.A., Haas, J., Shagina, L., Larson, E. and Friedman, C. Extracting Information on Pneumonia in Infants Using Natural Language Processing of Radiology Reports, *J Biomed Inform*, 2005, 38(4), 314–321.

[49] Pakhomov, S.V., Buntrock, J.D. and Chute, C.G., Automating the Assessment of Diagnosis Codes to Patient Encounters Using Example Based Machine Learning Techniques, *J Am Med Inform Assoc*, 2006, 13(5), 516–525.

[50] Crowley, R.S., Castine, M., Mitchell, K., Chavan, G., McSherry, T. and Feldman, M., caTIES: a Grid Based System for Coding and Retrieval of Surgical Pathology Reports and Tissue Specimens in Support of Translational Research, *J Am Med Inform Assoc*, 2010, 17(3), 253–264.

[51] Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B. and Duch, W., A Shared Task Involving Multi-label Classification of Clinical Free Text, *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, Prague, Czech Republic, 2007, 97–104.

[52] Open Health Natural Language Processing Consortium, OHNLP Documentation and Downloads, https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP_Documentation_and_Downloads [accessed May 31, 2010].

[53] IKITIK consortium, http://www.ikitik.fi, [accessed Jul 16, 2010].

[54] HEXAnord, http://dsv.su.se/en/research/ithealth/projects/hexanord, [accessed May 31, 2010].