Hanna Suominen

Machine Learning and Clinical Text

Supporting Health Information Flow

Turku Centre *for* Computer Science

TUCS Dissertations
No 125, December 2009

Machine Learning and Clinical Text

Supporting Health Information Flow

Hanna Suominen

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of the University of Turku, for public criticism in the lecture hall Beta in the ICT building on December 15, 2009, at 12 o'clock.

University of Turku
Turku Centre for Computer Science (TUCS)
Department of Information Technology
Joukahaisenkatu 3–5, 20520 Turku, Finland

2009

Supervisors

Tapio Salakoski
Professor, Vice Director
University of Turku, Department of Information Technology and TUCS
Turku, Finland

Helena Karsten
Research Director, Docent
Åbo Akademi University, Department of Information Technologies
Turku, Finland

Reviewers

Wray Buntine
Principal Researcher, Docent
National ICT Australia (NICTA, Australia's Information and Communications Technology (ICT) Centre of Excellence)
Canberra, Australia

Pierre Zweigenbaum
Senior Researcher, Associate Professor
LIMSI-CNRS (Computer Sciences Laboratory for Mechanics and Engineering Sciences, National Center for Scientific Research)
ERTIM-INALCO (Equipe de Recherche en "Textes, Informatique, Multilinguisme", Institut National des Langues et Civilisations Orientales)
Orsay and Paris, France

Opponent

Jussi Karlgren
Senior Researcher, Docent
Swedish Institute of Computer Science (SICS)
Kista, Sweden

ISBN 978-952-12-2375-4
ISSN 1239-1883

Abstract

Fluent health information flow is critical for clinical decision-making. However, a considerable part of this information is free-form text and inabilities to utilize it create risks to patient safety and cost-effective hospital administration. Methods for automated processing of clinical text are emerging.

The aim in this doctoral dissertation is to study machine learning and clinical text in order to support health information flow.

First, by analyzing the content of authentic patient records, the aim is to specify clinical needs in order to guide the development of machine learning applications. The contributions are a model of the ideal information flow, a model of the problems and challenges in reality, and a road map for the technology development.

Second, by developing applications for practical cases, the aim is to concretize ways to support health information flow. Altogether five machine learning applications for three practical cases are described: The first two applications are binary classification and regression related to the practical case of topic labeling and relevance ranking. The third and fourth application are supervised and unsupervised multi-class classification for the practical case of topic segmentation and labeling. These four applications are tested with Finnish intensive care patient records. The fifth application is multi-label classification for the practical task of diagnosis coding. It is tested with English radiology reports. The performance of all these applications is promising.

Third, the aim is to study how the quality of machine learning applications can be reliably evaluated. The associations between performance evaluation measures and methods are addressed, and a new hold-out method is introduced. This method contributes not only to processing time but also to the evaluation diversity and quality.

The main conclusion is that developing machine learning applications for text requires interdisciplinary, international collaboration. Practical cases are very different, and hence the development must begin from genuine user needs and domain expertise. The technological expertise must cover linguistics, machine learning, and information systems. Finally, the methods must be evaluated both statistically and through authentic user-feedback.

Acknowledgments

I owe my deepest gratitude to my supervisors Tapio Salakoski and Helena Karsten. You elevated my learning curve when growing as a researcher. You taught me crucial academic and managerial skills, gave me freedom and responsibility, while encouraging innovation and interdisciplinary skills. Thank you for demanding ambitiousness and believing in me.

My sincerest thanks go to Wray Buntine, Pierre Zweigenbaum, and Jussi Karlgren, my reviewers and opponent. The independent reviews were very constructive and complementary to each other. Wray, I greatly admire your vast knowledge of theoretical and applied machine learning and human language technology. Pierre, thank you for sharing your widely recognized expertise in human language technology and health informatics. Jussi, I acknowledge your research on human language technology and performance evaluation, and am looking forward to your feedback in my public defense.

I express my gratitude to my colleagues in the University of Turku at the Departments of Biostatistics, Information Technology, Mathematics, Nursing Science, and TUCS; in the Åbo Akademi University at the Department of Information Technologies; in the University of Tampere at the Institute of Medical Technology; in the Hospital District of Southwest Finland; and at Salivirta & Partners: Antti Airola, Barbro Back, Jorma Boberg, Jari Björne, Riitta Danielsson-Ojala, Minna Ervast, Jari Forsström, Filip Ginter, Satu Haapalainen-Suomi, Päivi Haltia, Anna Hammais, Katri Haverinen, Juho Heimonen, Hans Helenius, Ulrika Henriksson, Marketta Hiissa, Kai Kimppa, Elina Kontio, Heikki Korvenranta, Veronika Laippala, Tuija Lehtikunnas, Timo Leipälä, Shuhua Liu, Heljä Lundgrén-Laine, Dorina Marghescu, Eija Nordlund, Arja Pekonen, Tapio Pahikkala, Juha Perttilä, Sampo Pyysalo, Pentti Riikonen, Mia Ryhänen, Sanna Salanterä, Martti Tolvanen, Evgeni Tsivtsivadze, Leena Uronen, Mauno Vihinen, Riikka Vuokko, Matti Vuorinen, and others. You improved my performance and let me experience the power of successful team-work.

I am very grateful of having had the privilege of working with all extremely talented people in IKITIK, Louhi, and HEXAnord. Sari Ahonen, Juhani Reiman, Juhani Selänniemi, and Simo Vihjanen at Lingsoft; Peter Nyberg at Duodecim Medical Publications; and Mika Willberg at Acent-

ra: You have not only provided me state-of-the-art technology and linguistic resources but also tailored them. The Hercules Dalianis group in the DSV/KTH-Stockholm University; Richárd Farkas and György Szarvas in the University of Szeged; Peter Frank at ScanBalt; William Goossen at Results4Care; Janne Lahtiranta, Carita Martin, Tero Piispanen, and Carl-Johan Åkerblom at Turku Science Park; the Tarja Laitinen group in the University of Helsinki; Henning Müller in the University of Applied Sciences Western Switzerland; Klaus Oesch and Risto Wallin at Kites Association; and the Madis Tiik group at Estonian eHealth Foundation: Thanks for your national and international perspective.

I owe greatly to the TUCS community for four years funding and infrastructure. Christel Donner, Jouni Isoaho, Timo Järvi, Satu Jääskeläinen, Irmeli Laine, Tomi Mäntylä, Hannu Tenhunen, and others: I highly value your contribution. I am grateful to the Academy of Finland, Tekes – the Finnish Funding Agency for Technology and Innovation, and Turku Science Park for funding projects that have greatly supported this dissertation as well as studying commercialization of my research. I acknowledge travel funding from August and Lyydia Heino Foundation, Turku University Foundation, and Turku PET Centre as well as the computational resources of CSC – IT Center for Science. My heartfelt thanks go to the Business Development Laboratory of the Turku School of Economics, Finnish Society for Computer Science, organizers of the International Medical Natural Language Processing Challenge 2007, Student Award Committee of the 9th International Congress on Nursing Informatics, teachers of Loimaa Senior High, and organizers of Venture Cup 2007–2008 for encouraging my work.

I am ever so grateful for the Australian National University, Research School of Information Sciences and Engineering as well as NICTA, Canberra and Queensland Research Laboratories. Marcus Hutter, Robert C Williamson, Marconi Barbosa, Terry Caelli, Leif Hanlen, Penelope Sanderson, and others: you most heartily welcomed me to your community and helped me recognize directions for my post-doctoral research.

I express my greatest gratitude to my family and friends in Finland and abroad for your support. Anne, Anu, Maria, Minna, Piia, all wonderful people at Raija Lehmussaari Ballet School and Bailes Cubanos, as well as my other close friends and family members: I would be very lonely, weak, and sick without you. I acknowledge Raija for showing me most of what I know about teaching. Jeanette Strole Parks, my dear cousin, I am most grateful to you for thoughtful proof-reading. You also spurred this ex-cheerleader on to finish her dissertation. Thank you, my parents, grandparents, Matti, and Tuuli for always being tower of strength to me. You have taught me the importance of science, education, and commitment. You truly know me — my naked self with all my flaws, temper, and crazy working hours — and still give me your unconditional love and care.

To my dear family

List of Original Publications Included in the Dissertation

Paper I. Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2007). Applying language technology to nursing documents: pros and cons with a focus on ethics. *International Journal of Medical Informatics*, 76(S2):S293–S301.

Extends:
Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2006). Theoretical considerations of ethics in text mining of nursing documents. In Park, H.-A., Murray, P., and Delaney, C., editors. *Consumer-Centered Computer-Supported Care for Healthy People. Proceedings of NI2006, the 9th International Congress of Nursing Informatics*, volume 122 of *Studies in Health Technology and Informatics*, pages 353–364. IOS Press, Amsterdam, the Netherlands.
(Student encouragement award)

Paper II. Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2007). Towards automated classification of intensive care nursing narratives. *International Journal of Medical Informatics*, 76(S3):S362–S368.

Extends:
Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2006). Towards automated classification of intensive care nursing narratives. In Hasman, A., Haux, R., Van Der Lei, J., De Clercq E., and Roger France, F. H., editors. *Ubiquity: Technologies for Better Health in Aging Societies. Proceedings of MIE 2006, the 20th International Congress of the European Federation of Medical Informatics*, volume 124 of *Studies in Health Technology and Informatics*, pages 789–794. IOS Press, Amsterdam, the Netherlands.

Paper III. Suominen, H., Pahikkala, T., Hiissa, M., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2006). Relevance ranking of intensive care nursing narratives. In Gabrys, B., Howlett, R. J., and Jain, L. C., editors. *Knowledge-Based Intelligent Information and Engineering Systems. Proceedings of the 10th International Conference, KES 2006, Part I*, volume 4251 of *Lecture Notes in Computer Science*, pages 720–727. Springer, Berlin / Heidelberg, Germany.

Paper IV. Ginter, F., Suominen, H., Pyysalo, S., Salakoski, T. (2009). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application. *International Journal of Medical Informatics*, 78(12):e1–e6.

Extends:
Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2008). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application. In Salakoski, T., Rebholz-Schuhmann, D., and Pyysalo, S., editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, SMBM 2008*, number 51 of *TUCS General Publication*, pages 37–44. Turku Centre for Computer Science, Turku, Finland.
and
Suominen, H., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). Automated text segmentation and topic labeling of clinical narratives. In Karsten, H., Back, B., Salakoski, T., Salanterä, S., and Suominen, H., editors. *Proceedings of the First Conference on Text and Data Mining of Clinical Documents, Louhi 2008*, number 52 of *TUCS General Publication*, pages 99–103. Turku Centre for Computer Science, Turku, Finland.

Paper V. Suominen, H., Ginter, F., Pyysalo, S., Airola, A., Pahikkala, T., Salanterä, S., and Salakoski, T. (2008). Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In Hauskrecht, M., Schuurmans, D., and Szepesvari, C., editors. *Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications*, 2008 July 9, Helsinki, Finland.

Paper VI. Suominen, H., Pahikkala, T., and Salakoski, T. (2008). Critical points in assessing learning performance via cross-validation. In Honkela, T., Pöllä, M., Paukkeri, M.-S., Simula, O., editors. *Proceedings of the 2nd International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR 2008*, pages 9–22. Multiprint, Espoo, Finland.

Paper VII. Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2009). Efficient cross-validation algorithms for sparse regularized least-squares. Unpublished manuscript.

Extends:
Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2009). Efficient hold-out for subset of regressors. In Kolehmainen, M., editor. *Adaptive and Natural Computing Algorithms, 9th International Conference, ICANNGA 2009, Revised Selected Papers*, volume 5495 of *Lecture Notes in Computer Science*, pages 350–359. Springer, Berlin / Heidelberg, Germany.

List of Related Publications not Included in the Dissertation

Co-authored publications

International per-reviewed book chapters

- Suominen, H., Pyysalo, S., Hiissa, M., Ginter, F., Liu, S., Marghescu, D., Pahikkala, T., Back, B., Karsten, H., and Salakoski, T. (2008). Performance evaluation measures for text mining. In Song, M. and Wu, Y.-F. B., editors. *Handbook of Research on Text and Web Mining Technologies*, volume 2, pages 724–747. IGI Global, Hershey, Pennsylvania, USA.

Editorials in international per-reviewed journals

- Suominen, H. and Karsten, H. (2009). Mining of clinical and biomedical text and data: editorial of the special issue. *International Journal of Medical Informatics*, 78(12):786–787.

Papers in domestic per-reviewed journals

- Suominen, H. (2008). Kieliteknologiasta ratkaisu potilasdokumentaation hyödynnettävyyteen [Human language technology as a solution for usability of patient records]. *Tietojenkäsittelytiede* [*Computer Science*], 28:4–14.

Papers in international per-reviewed conference proceedings

- Lundgrén-Laine, H., Suominen, H., Kontio, E., and Salanterä, S. (2009). Intensive care admission and discharge — critical decision-making points. In Saranto, K., Flatley Brennan, P., Park, H.-A., Tallberg, M., and Ensio, A., editors. *Connecting Health and Humans.*

Proceedings of NI2009, the 10th International Congress of Nursing Informatics, volume 146 of *Studies in Health Technology and Informatics*, pages 358–361. IOS Press, Amsterdam, the Netherlands.

- Pahikkala, T., Airola, A., Suominen, H., Boberg, J., and Salakoski, T. (2008). Efficient AUC maximization with regularized least-squares. In Holst, A., Kreuger, P., and Funk, P., editors. *Proceedings of the Tenth Scandinavian Conference on Artificial Intelligence, SCAI 2008*, volume 173 of *Frontiers in Artificial Intelligence and Applications*, pages 12–19. IOS Press, Amsterdam, the Netherlands.

- Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2007). Transductive ranking via pairwise regularized least-squares. In Frasconi, P., Kersting, K., and Tsuda, K., editors. *The 5th International Workshop on Mining and Learning with Graphs, MLG 2007*, pages 175–178. 2007 August 1–3, Florence, Italy.

- Suominen, H. J., Lehtikunnas, T., Hiissa, M., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2005). Natural language processing for nursing documentation. In Fonseca, J. M., editor. *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare, CIMED 2005*, pages 147–154. 2005 June 29 – July 1, Costa da Caparica, Portugal.

- Suominen, H., Lundgrén-Laine, H., Salanterä, S., Karsten, H., and Salakoski, T. (2009). Information flow in intensive care narratives. In Chen, J., Chen, C., Ely, J., Hakkani-Tr, D., He, J., Hsu, H.-H., Liao, L., Liu, C., Pop, M., Ranganathan, S., Reddy, C.K., Ruan, J., Song, Y., Tseng, V.S., Ungar, L., Wu, D., Wu, Z., Xu, K., Yu, H., Zelikovsky, A., editors. *Proceedings IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBM 2009*, pages 325–330. Institute of Electrical and Electronics Engineers, Los Alamitos, California, USA.

- Suominen, H., Lundgrén-Laine, H., Salanterä, S., and Salakoski, T. (2009). Evaluating pain in intensive care. In Saranto, K., Flatley Brennan, P., Park, H.-A., Tallberg, M., and Ensio, A., editors. *Connecting Health and Humans. Proceedings of NI2009, the 10th International Congress of Nursing Informatics*, volume 146 of *Studies in Health Technology and Informatics*, pages 191–196. IOS Press, Amsterdam, the Netherlands.

- Suominen, H. and Salakoski, T. (2009). Information retrieval and personal health records: user needs and domain tailoring. In *SPIRE*

2009 Workshop on Task-based Information Access. 2009 August 28, Saariselkä, Finland. In Press.

Papers in domestic per-reviewed conference proceedings

- Hiissa, M., Suominen, H., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2006). Kohti tehohoitotyön narratiivien tehokkaampaa hyödyntämistä luonnollisen kielen käsittelyn avulla [Towards more efficient use of intensive care narratives through natural language processing]. In Häyrinen, K., editor. *Sosiaali- ja terveydenhuollon tietotekniikan ja tiedonhallinnan tutkimuspäivät* [*Research Convention for Information Technology and Data Management in Welfare and Health*], volume 18/2006 of *Työpapereita, Tutkimuspaperit 2006, the Centre for Research and Development of Welfare and Health, Stakes*, pages 17–23. Valopaino Oy, Helsinki, Finland.

- Suominen, H. (2007). Kieliteknologian menetelmien soveltaminen potilasdokumentaation hyödyntämiseen [Applying human language technology for patient documents]. In Koskinen, M. and Jauhiainen, E., editors. *Tietojenkäsittelytieteen päivät 2007* [*The Computer Science Convention 2007*], volume TU-25 of *Tutkimuksia, Tietojenkäsittelytieteiden julkaisuja*, pages 46–50. Jyväskylä University Press, Jyväskylä, Finland.

Papers in domestic professional journals

- Lundgrén-Laine, H. and Suominen, H. (2009). Mistä näitä tietoja oikein tulee? Kieliteknologialla tehohoidon tiedot hallintaan [Where do all these data come from? Solving problems in knowledge management of intensive care data via human language technology]. *Tehohoito* [*Journal of the Finnish Society of Intensive Care, FSIC*], 27(2):102–104.

Abstracts in domestic per-reviewed conference proceedings

- Lundgrén-Laine, H., Hiissa, M., Suominen, H., Lehtikunnas, T., and Salanterä, S. (2006). Tehohoidon elektroninen kirjaaminen — kohti narratiivien automaattista luokittelua [Electronic intensive care documents — towards automated classification of narratives]. In Salanterä, S., Hupli, M., and Axelin, A., editors. *IX Kansallinen hoitotieteellinen konferenssi "Viisas vaikuttaja"* [*IX National Nursing Science Conference "Wise opinion leader"*], page 98. University of Turku, Turku, Finland.

- Suominen, H., Lundgrén-Laine, H., Perttilä, J., Salakoski, T., and Salanterä, S. (2008). Tehohoidon elektroniset potilasasiakirjat — hyödyntämätön voimavara [Electronic intensive care patient records — an untapped resource]. In *The Finnish Medical Convention*. 2008 January 6–10, Helsinki, Finland.

Co-edited conference proceedings

- Karsten, H., Back, B., Salakoski, T., Salanterä, S., and Suominen, H. editors (2008). *Proceedings of the First Conference on Text and Data Mining of Clinical Documents, Louhi 2008, Turku, Finland*, number 52 of *TUCS General Publication*. Turku Centre for Computer Science, Turku, Finland.

Contents

List of Figures

List of Tables

List of Acronyms

HLT	—	human language technology
HMM	—	hidden Markov model
ICD-10	—	International Classification of Diseases, 10th Revision
ICD-9-CM	—	International Classification of Diseases, 9th Revision, Clinical Modification
ICP	—	intracranial pressure
ICU	—	intensive care unit
LSA	—	latent semantic analysis
LOS	—	length of stay
RLS	—	regularized least-squares
ROC	—	receiver operating characteristic
UMLS	—	Unified Medical Language System

Part I

Research Summary

Chapter 1

Introduction

This doctoral dissertation studies machine learning and clinical text in order to support health information flow. Because of emphasized demands on reliable solutions in health care, special care is shown to performance evaluation. This discussion is broadened from clinical applications to machine learning theory in general.

1.1 Clinical text and human language technology

Efficient access to the gathered *health information* is critical for accurate and timely decision making in *clinical settings* (i.e., includes direct observation of the patient), or more generally, in health care. Health information refers to "any information, whether oral or recorded in any form or medium, that

a) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse, and

b) relates to the past, present, or future physical or mental health or condition of an individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual"

(US Department of Health & Human Services, 1996). By using *electronic health information systems*, millions of gigabytes of health information are generated annually as *electronic health records* (or *electronic patient records* in hospitals), and these volumes of data are significantly larger than in many other domains (Cios and Moore, 2002).

For example, every year in Finland with a population of circa five million, about thirteen million general practitioner or occupational health-care physician visits, almost seven million specialized care outpatient visits, and

Photos: University of Turku, University Communications

Figure 1.1: *Electronic health records and health communication in practice*

over a million specialized health care periods take place (Stakes, 2008, pp. 105, 116). Each of these events is documented in health records (Decree 99/2001 of the Ministry of Social and Health concerning documenting and storing patient information), and because of increasing storing capabilities in the electronic form, larger and larger quantities of content is gathered and stored.

At the same time, the content of health records has became increasingly complex and unintelligible. A considerable part of health records is free-form written or spoken text, *narratives* (Figure 1.1). Narratives are an easy, nuanced and natural way of expression. In health care, they are often a necessity for a thorough and precise explanation of a given event. However, this expressive power inherently bears a substantial ambiguity and personal differences in vocabulary and style (Lovis et al., 2000). This complicates the use of the gathered information.

Information search from narratives is often difficult, laborious, and time consuming. Currently, narrative information is severely underutilized in clinical judgment and decision making. Electronic health information systems have been shown, instead of freeing up time for direct care, to lead health care professionals spending more time on laborious, unproductive

data manipulation and formatting (Pierpont and Thilgen, 1995; Smith et al., 2005). In addition, evidence has been given that these systems support information gathering but not its active utilization (Snyder-Halpern et al., 2001). Pointedly, from the point of view of narrative information, electronic health records are currently like write-only memories.

As a remedy for the current situation, standardization and structuring of narratives have been proposed. Indeed, a standard, more compact, and unambiguous language can ease *information access* (i.e., satisfying "human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language" (Allan et al., 2003)) and improve search results. However, this approach has also some problems.

First, substituting numerical and structured data for narratives is laborious and leads easily to differences and errors in the coding. Nurses use on average from 12 percent to 35 percent of their working time on electronic care documentation (Hakes and Whittington, 2008; Banner and Olney, 2009). If the narratives are structured and typed manually, nurses' time available for other duties will only decrease: According to the analysis of health records related to approximately $5,600$ admissions and $41,000$ complete patient-days, the number of items that nurses type into the records has increased by 26 percent from circa $1,200$ items per patient-day in 2000 to circa $1,500$ items per patient-day in 2005 (Manor-Shulman et al., 2008). This does not cover narrative parts of the records. The amount of narratives that nurses write is substantial: the patient-wise average is over 11 pages during an intensive care in-patient period alone (see Section 2.5). Moreover, if the standardization and structuring are performed manually, text ambiguity and personal differences are bound to cause inconsistencies.

Second, dispensing with narratives against structured data may lead to a significant information loss because of limiting the expressive power (Lovis et al., 2000; Walsh, 2004). In addition, this approach offers weaker support to individualized care (Tange, 1996; Tange et al., 1997).

In conclusion, increased standardization and structuring contributes to information access. However, research should aim to develop automated ways that achieve this outcome more systematically and retain the benefits of narratives, that is, the expressive power and ease of production.

Tools for automated processing of narratives, *human language technology* (HLT) can be used to support *information flow* in patient narratives (Figure 1.2). This covers both producing and using information; information flow is defined as "links, channels, contact, flow of communication to pertinent people or groups in the organization" (Glaser et al., 1987). The HLT approach is increasingly gaining the interest of both health care practitioners and academic researchers. It allows supporting not only health care professionals' work, but individuals (i.e., patients or customers of health

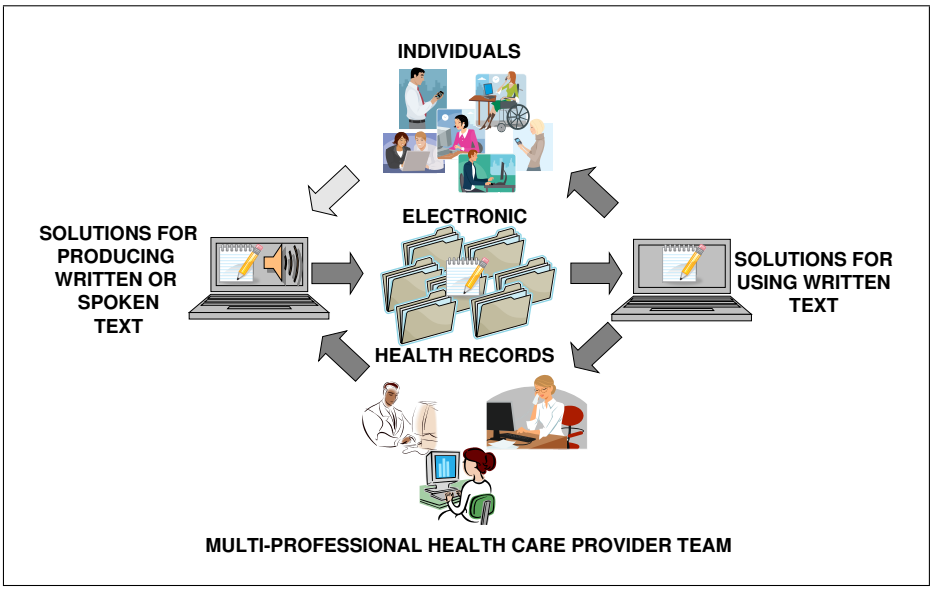


Figure 1.2: *Human language technology for producing and using health records*

care services, or citizens) could also use the technology themselves to ease in accessing their health data, and in the future, even enter data to their records.

HLT refers to computational tools that analyse and generate natural, human language. These analysis and generation techniques compose the field of *natural language processing*. A closely related field is *text mining*, or text data mining; typical tasks in both natural language processing and text mining are

- classifying textual documents according to the similarity of their contents,
- retrieving documents whose content matches the query, and ranking the output list in the relevance order,
- extracting information about specific predetermined topics, and
- summarizing the document content.

However, to make a distinction, text mining refers to analyzing a large collection of narratives and producing an output that can also be inferred, and not explicitly stated in the input. Because this dissertation does not focus on large data collections and an inferred output only, the broader term of natural language processing is more adequate. Moreover, because the

overall goal is to concretize the development of real-life health applications, I prefer using the term HLT.

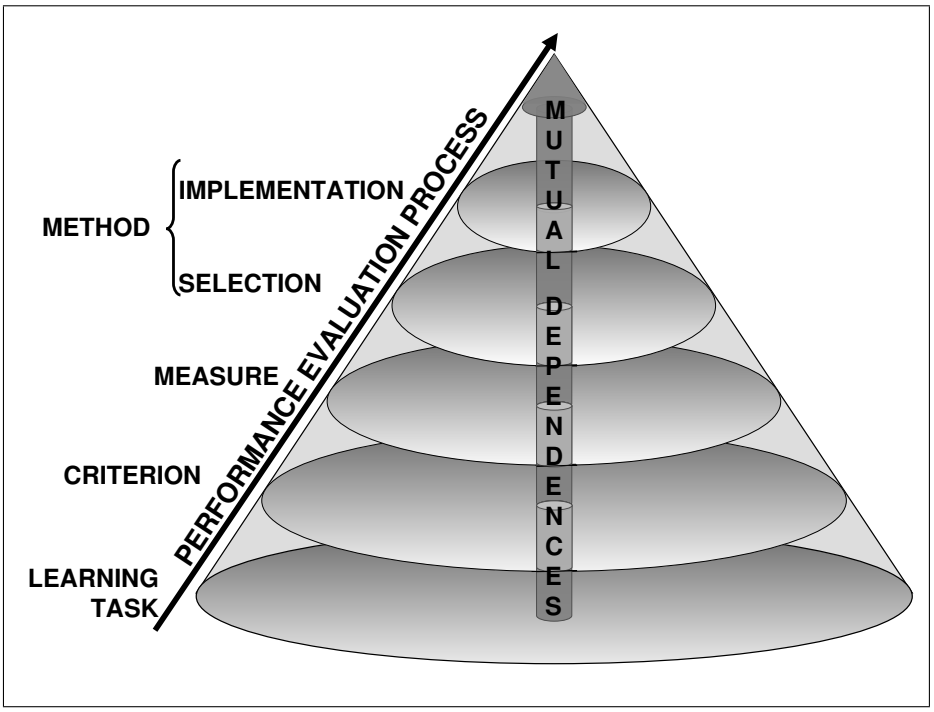
Tailoring is needed in order to obtain HLT solutions for highly specialized language in electronic health records. The records contain specific professional jargon, which limits the linguistic coverage of tools for general languages (Huang et al., 2005; Laippala et al., 2009). The earliest health-related entries on HLT are from the early 1970s[1], and as pioneering studies Becker (1972); Anderson et al. (1975); Hirschman et al. (1976); Collen (1978); Young (1982); Chi et al. (1983); Shapiro (1983); as well as Gabrieli and Speth (1986) can be mentioned. This PubMed search returns almost 2,200 references which shows that HLT is gaining more and more interest. However, the successful application of systems developed for one language to another is not straightforward (see, e.g., from English to Portuguese (Castilla et al., 2007) and from English to Swedish (Velupillai et al., 2009)). At the moment, HLT is rarely used in electronic health records, in particular with Finnish and many other European languages, although promising results have been accomplished in very specific tasks, as this dissertation will illustrate. More work is needed to connect these components and actors together to develop more comprehensive solutions that establish themselves in electronic health records.

1.2 Machine learning and performance evaluation

HLT aims to understand textual content, which requires not only linguistic techniques but also *machine learning*. Machine learning, with its origin in computer science, mathematics, and statistics, is a field with a focus on techniques that enable machines (i.e., computers) to learn to carry out specific tasks by following a certain automated learning algorithm. Traditionally, the techniques are divided into *supervised*, *unsupervised*, and *semi-supervised* learning. In supervised learning, the task is to learn a mapping that connects an input with its desired output. In other words, a data set defining the desired outputs called *gold standard* (aka reference standard) is known in advance and exploited in learning. In contrast, in unsupervised learning, the learning algorithm is given the inputs without the desired outputs. Finally in semi-supervised learning, some inputs are associated with the desired outputs to be used in learning.

Performance evaluation assesses learning and generalization capabilities of machine learning applications and algorithms, and it is used for *training*, *model selection*, and *testing*. With supervised techniques, training refers to the process of selecting among a set of candidates the mapping that best

[1]Hanna Suominen conducted a PubMed search on 2009 November 21. It utilized the MeSH hierarchy with the query "(natural language processing OR (text mining)."



Adapted from Paper VI

Figure 1.3: *Performance evaluation as a process*

performs in the task in question. The data used for this purpose is called a *training set*. The mappings often include parameters, or the learning is based on some other kind of parametrization. Model selection (a.k.a. validation) refers to fixing the values of the parameters by evaluating performance with a spectrum of various values and choosing the best. The data used for this purpose is called a *validation set*. Finally, testing assesses the final performance and the related data is called a *testing set*.

Performance evaluation can be seen as a cascaded process (Figure 1.3, Spärck Jones and Galliers (1996, pp. 19, 20), Hirschman and Thompson (1997), Paper VI): The first step is to specify the *learning task* (e.g., to identify the patients with a given diagnosis from a collection of narratives). The second is to define the *evaluation criterion* (e.g., the correctness of diagnosis coding output). The third is to select an *evaluation measure* that describes this criterion (e.g., the proportion of correct diagnoses to all diagnosing decisions). The fourth — and the last — step consists of choosing and implementing an *evaluation method* that determines a value for the measure by specifying the use of data.

Hold-out techniques constitute one of the most popular groups of performance evaluation methods. They are based on setting a part of the data aside. For example, the data can be divided into disjoint sets so that a third is used for training, a third for model selection, and a third for testing. In particular, *N-fold cross-validation* and *leave-one-out cross-validation* are widely used. In N-fold cross-validation, data is divided into N approximately equal sets. Each set is used in turn for testing while the rest are used for training and model selection. During each repetition, the performance is evaluated by using the chosen measure. The final value of the measure is obtained by averaging the the N values. Leave-one-out cross-validation refers to the special case of N-fold cross-validation, where one data point in turn is held out for testing.

Getting reliable results from the performance evaluation is difficult. The steps of the performance evaluation process are mutually dependent and each of them includes a number of options to choose from (Dietterich (1998), Bengio and Grandvalet (2004), Paper VI). The choices and their implementation must be made carefully in order to develop high-quality machine learning applications and algorithms. Further, the dependences between the evaluation steps are poorly understood. Finally, the more complex outputs and hybrids of machine learning components are needed, the more difficult the evaluation gets.

1.3 Research objectives

The aim of this doctoral dissertation is to study and concretize the application of machine learning to processing textual health records (Table 1.1). My research questions are

1. Why is HLT needed in the clinical practice?
    a) What is the ideal information flow like?
    b) What are the problems and challenges in reality?
    c) Which tasks can machine learning support?
2. What are the methods to build this support?
    a) Which existing machine learning methods can be applied?
    b) What kind of method development is needed?
    c) What is the achieved performance in terms of statistical metrics?
3. How is quality evaluated reliably?
    a) How should the evaluation measure be selected?

b) How should the evaluation method be selected?

c) How should the evaluation method be implemented?

The first category of questions addresses clinical information needs. Its purpose is to ensure clinical significance of the study, and a deeper understanding of the data in order to lay groundwork for the HLT development. Of the three question categories, it is closest to health and information systems sciences. Based on these clinical needs, the second question category focuses on evaluating and developing machine learning applications. This applied machine learning research goes hand in hand with the basic machine learning method development, and motivates the third question category. The third category broadens the discussion from clinical applications to machine learning theory in general by addressing reliable performance evaluation. Additional motivation for performance evaluation comes from the emphasized quality requirements for clinical HLT because of care safety. The iterative phase of clinical trials and further method development is outside the scope of this dissertation.

In comparison with Figure 1.2, my focus is on supporting health care professionals in the use of gathered narratives. All data is in written form, and it has been produced by nurses, physicians, and radiologists. The main domain is Finnish *intensive care units* (ICUs). This is complemented with HLT for automated analysis of English narratives from a US radiology department. ICUs have been chosen for two reasons: First, its complexity, information intensity, multi-professionalism, and fast pace due to critically ill patients, emphasize fluent information flow and increase needs for decision support. Second, the information needs of ICU professionals can be assumed to be internationally similar. Therefore, HLT supporting decision making should be internationally applicable. Based on a comparison of Canadian, Finnish, Northern Irish, Swiss, and US health care professionals working in intensive care, psychiatric care, public health, and short and longterm care, decision making in ICUs is internationally most similar (Lauri and Salanterä, 2002).

Table 1.1: *Overview of the dissertation*

Topics	Research questions	Sections	Papers
Introduction			
Clinical text and human language technology		1.1	
Machine learning and performance evaluation		1.2	
Research objectives		1.3	
Contributions		1.4	
Organization of the dissertation		1.5	
Clinical needs			
Ideal information flow	1a	2.2	
Problems and challenges in reality	1b	2.5	
Clinical needs for human language technology	1c	2.6	I
Machine learning applications and their performance			
Topic labeling and relevance ranking (2 applications)	2a, 2b, 3a	3.1	II, III
Topic segmentation and labeling (2 applications)	2a–c, 3a	3.2	IV
Diagnosis coding (1 application)	2a–c, 3a	3.3	V
Evaluation reliability			
Evaluation measures	3a	3.1.3, 3.2.3, 3.3.3	V, VI
Evaluation methods and their implementation	3b	4	VI, VII
Conclusions, significance, and future work			
Conclusions		5.1	
Computer science: significance and future work		5.2	
Health care: significance and further work		5.3	

1.4 Contributions

The contributions of the dissertation are summarized as follows:

1. Clinical needs
    a) Model of the ideal information flow
    b) Model of the problems and challenges in reality
    c) Road map for the HLT development
2. Machine learning applications and their performance
    a) Practical case of topic labeling and relevance ranking
        - Two applications: binary classification and regression
        - Three topics: breathing, blood circulation, pain
    b) Practical case of topic segmentation and labeling
        - Two applications: supervised and unsupervised multi-class classification
        - Six topics: breathing, hemodynamics, consciousness, relatives, diuresis, other
    c) Practical case of diagnosis coding
        - One application: multi-label classification
        - 45 diagnosis codes
3. Evaluation reliability
    a) Associations between evaluation measures related to the five clinical applications
    b) Associations between hold-out methods and their implementation
    c) A new hold-out method for a particular learning algorithm

1.5 Organization of the dissertation

This research summary is designed to stand alone as an independent entity, where the ideas and methods introduced in Papers I–VII have been connected together and some new results are introduced. The learning methods are elaborated in a detail in the referred papers. (See Table 1.1.) The dissertation is organized according to the three main research questions into the chapters of Clinical needs (Chapter 2), Machine Learning Applications (Chapter 3), and Performance Evaluation Methods and Evaluation Reliability (Chapter 4). Performance evaluation measures are discussed separately for each machine learning application in Chapter 3.

Chapter 2 analyses the clinical HLT needs in Finnish intensive care by addressing both the ideal information flow and reality with narratives. The outcome is a road map that combines the ideal information flow with future HLT components. Because the results are mainly previously unpublished, the discussion is relatively detailed.

In accordance with the road map, Chapter 3 presents five applications of machine learning to support the flow. Four applications ease information access through automated relevance evaluation between a given search topic and text segments: the first is for binary classification to topically relevant and irrelevant segments, the second for gradual relevance ranking, the third for automatically dividing text into topical segments and simultaneously assigning the topic labels in a supervised manner, and the fourth application performs the topic segmentation and labeling task in an unsupervised manner. These four applications are tested with Finnish ICU *nursing narratives*, that is, narratives written by nurses to describe nursing care. The fifth application is for diagnosis coding, and it is tested with English narratives written by physicians and radiologists.

Chapter 4 discusses performance evaluation methods for machine learning algorithms and applications. It takes a particular learning algorithm as an example and introduces an evaluation method for this algorithm. This hold-out method is computationally efficient, contributes to evaluation reliability, and allows holding out multiple inputs simultaneously. The method is described in the referred paper.

Finally, Chapter 5 concludes the dissertation, discusses its significance, and presents topics for future work.

Chapter 2

Clinical Needs

This chapter specifies clinical needs for HLT in intensive care. First, as authentic health records are a prerequisite for this thesis, legal and ethical aspects are discussed. For a more thorough and detailed discussion, see Suominen et al. (2006) and Paper I.

Then, the chapter describes an ideal ICU information flow and addresses its current problems and challenges via data-driven analysis. The results are mainly previously unpublished: Preliminary results have been described in Suominen et al. (2009a), the chapter extends it with an emphasized focus on ICU information flow as a whole.

I have conducted the extension with authors of Suominen et al. (2009a) as well as with Jari Forsström (Salivirta & Partners, Finland) and Juha Perttilä (Turku University Hospital, Adult Intensive Care Unit, Finland). My role has been the first author. Together with Lundgrén-Laine, I have been responsible for the study conception, design, data analysis, and drafting of the manuscript. Lundgrén-Laine has performed the data collection and provided ICU documentation and decision making expertise. I have automated content analysis and performed statistical analyses. Salanterä, Forsström, Perttilä, Karsten, and Salakoski have critically commented on the study, and Salanterä, Karsten, and Salakoski have supervised it.

Based on this data-driven analysis, the chapter then combines the detected needs with the ideal model of the information flow in a road map for developing HLT. It considers the care documentation and text use process in accordance with the ideal ICU information flow model. Its aim is to focus the development in order to create comprehensive HLT solutions by meeting the user needs. Finally, the chapter compares the road map with related work.

The road map is previously unpublished. Its initial version without the proof-reading component is described in our abstract and poster Suominen et al. (2008c), and paper Lundgrén-Laine and Suominen (2009). We have

discussed potential HLT components and their benefits in Suominen et al. (2005), Paper I, Suominen et al. (2006), and Lundgrén-Laine et al. (2009). However, these papers do not combine the components in any way.

2.1 Legal and ethical aspects

Authentic health records are a prerequisite for this thesis; without them developing intelligent, domain-tailored HLT solutions and evaluating their quality is impossible. In total, two confidential sets of Finnish health records have been used in this thesis (Details of their content and collection are addressed in Sections 2.3 and 3.1.5). Storing and using this data in HLT development requires careful consideration and compliance with legal and ethical principles.

The Finnish legislation related to personal data or health care does not address HLT as a special case. In most instances, when researchers and tool developers access electronic health records for study purposes, they follow the security procedures designed for accessing electronic health records for health care purposes (Berman, 2002). For example, when using Finnish electronic health records, researchers must conform to general Finnish laws, such as the Act on the Status and Rights of Patients 785/1992; Medical Research Act 488/1999; Personal Data Act 523/1999; Medical Research Decree 986/1999; and Decree 99/2001 of the Ministry of Social and Health concerning documenting and storing patient information. As an exception to the general requirement of the patient's consent (Statutes of Finland, Act on the Status and Rights of Patients 785/1992), Statutes of Finland, Personal Data Act 523/1999 declare that health records may be processed for the purposes of scientific research if

1. "the research cannot be carried out without data identifying the person and the consent of the data subjects cannot be obtained owing to the quantity of the data, their age or another comparable reason,

2. the use of the personal data file is based on an appropriate research plan and a person or a group of persons responsible for the research have been designated,

3. the personal data file is used and data are disclosed therefrom only for purposes of historical or scientific research and the procedure followed is also otherwise such that the data pertaining to a given individual are not disclosed to outsiders, and

4. after the personal data are no longer required for the research or for the verification of the results achieved, the personal data file is destroyed

> or transferred into an archive, or the data in it are altered so that the data subjects can no longer be identified."

In this thesis, I have adhered to the following practice in order to ensure patient confidentiality:

Permits. Owing to the quantity of the data needed in machine learning development, health records have been collected without patients' consent. Instead, we have furnished favorable opinions from ethical committees (Hospital District of South Karelia, record number 8/01; Hospital District of Southwest Finland, record number 401/2005; Jorvi, record number 11.1.2001; Kainuu Central Hospital, record number 1/2001; Lapland Hospital District, record number 1.2.2001; Satakunta Central Hospital, record number 7.2.2001) and proper permits from the respective hospital authorities having the right to acknowledge the permissions. The legal research officer of the Hospital District of Southwest Finland has also been consulted for assuring good research practice.

Research material. Before receiving the research material, all patient names and other similar pieces of personal information have been replaced with anonymous identification numbers. The researchers do not have access to de-identification keys. The only exceptions are unstructured parts of narratives, whose content cannot be fully de-identified automatically. All parties with permission to access the research material conform to the Finnish legislation, and have signed a written vow of silence as well as a contract specifying the required data protection actions. The material is stored behind passwords and used only for HLT development. Results are published in a way that guarantees the anonymity of individual patients. This anonymity is also assured, when providing material (i.e., examples illustrating common lexical and linguistic features) needed to tailor commercial HLT (e.g., Lingsoft FinTWOL Finnish Morphological Analyzer) to the research domain. When the study ends, the research material will be returned to its original hospital. There the data will be filed for possible future use. The hard drives, where researchers stored the data will be destroyed.

Risks. Research ethics and risks are carefully considered. We have studied them in detail in Suominen et al. (2006) and Paper I. However, despite carefully conforming to the legislation and good scientific practice (National Advisory Board on Research Ethics, 2002), the risk of violating patient confidentiality in HLT development still remains. The main risks in this thesis arise from the unstructured parts of narratives that have not been de-identified and from a possibility to deduce the patient identity by combining the gender, date, location and other similar details. However, no data is used in this thesis to recognize the identity of individual people and special care has been paid to data protection.

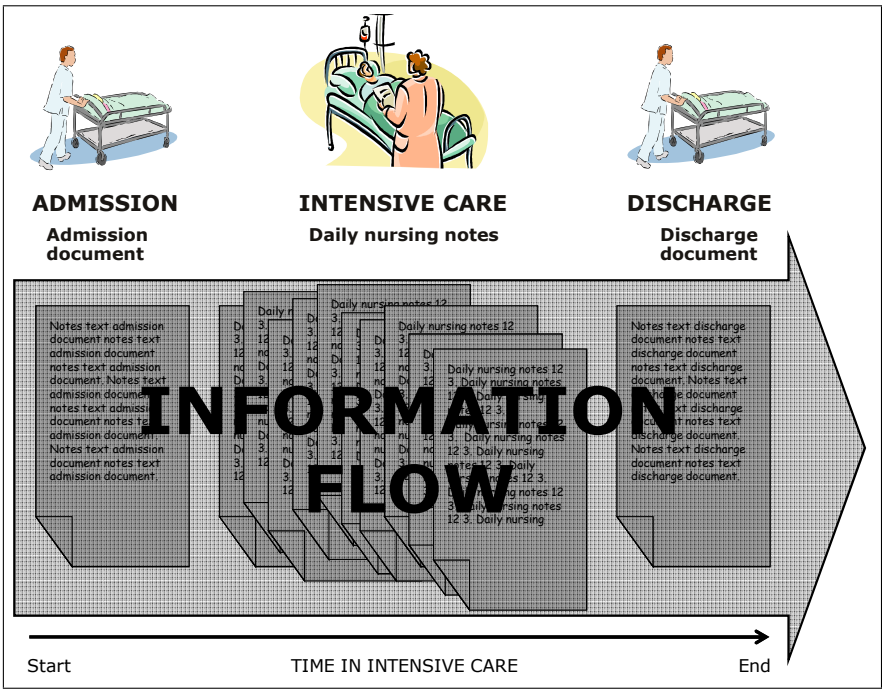


Figure 2.1: *Ideal information flow*

2.2 Ideal information flow in intensive care

The need for a fluent information flow is emphasized in intensive care. ICUs are highly specialized hospital wards delivering care for acutely and critically ill patients, who require continuous monitoring and observation of vital signs (Ambrosius et al., 1997; Society of Critical Care Medicine, 2007). As a result, a vast amount of data, both as narratives and in a numerical or structured format, is accumulated during care, and they are used to support clinical decision-making.

For high care quality it is necessary to get a picture of what is relevant in health records. That is, the patient information should flow fluently between the crucial decision-making points of admission and discharge (Lundgrén-Laine et al. (2009), Figure 2.1). Because the emerging HLT methods as well as computational and linguistic resources are enabling the development of tailored methods for clinical narratives (see Chapter 1 and Section2.6), my focus is on narratives. Electronic patient records in ICUs already offer numerous decision support tools based on processing numerical and structured data (Hanson and Marshall, 2001; Weaver et al., 2005).

ADMISSION DOCUMENT
PREVIOUS/OTHER DISEASES:
¤ *RR disease*
¤ *Chr. FA*
ANAMNESIS:
18.3 heavy chest pain started. In Loimaa, strong ST changes. Angiography and discovery of a tight main stenosis. LAD good, substantial changes in RCA and LCX. Tumefaction and intubation in the morning of 19.3 in TUH. DRUGNAME –infusion started. ICU ad. for the emergency operation.
O room:
Before perfusion, the patient has low pressures. A balloon pump will be placed, DRUGNAME going in. Pulmonal pressures high -> DRUGNAME and DRUGNAME. At the end of perfusion had to start DRUGNAMEinfusion also. Got 4 x ice plasma and don. thrombocytes.
Dg: MCC
Oprtn: Reconstr. coron. cordis No. IV
¤ *Ao -> DRUGNAME 140ml/h*
¤ *Ao -> DRUGNAME 103 ml/h*
¤ *Ao -> DRUGNAME 154 ml/h*

DAILY NURSING NOTES
…

Long night s.
RR tryed st. easily rise even to very high values towards midnight, towards morning RR level went down and became steady. p slightly tachycardic. Profuse diuresis.
Towards midnight fillpressures occ. highish, towrads morning went down subst.
Suff CI (c. level 2), rised ad 2.5 at morning.
With 40 % vm .oxidation on the low side, ventilates well. Put 50 % mask, which outcomes 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 pat., is opening eyes occ. and tryes to answer to the posed questions but lacks the strength often.

long shift
hemod: RR still mainly high and pulse tachy. Pulse occasionally sinus rhythm. Got 2 ampoules of DRUGNAME every hour before noon; this steadied pulse only slightly. In the afternoon, after DRUGNAME RR down very strongly.Pulmonal cannula rmved
breathing: in the thx scan increased pleural fluid; continued the patient's heavy "drying." Oxygenation improved during the day
diuresis been very profuse, DRUGNAME cont with the previous dose because of the lung situation
conciousssnes: 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 finger tubes and pass away. Got DRUGNAME 5 mg iv., which decreased the pressure substly and also made verytired. Also after DRUGNAME is extr. exhausted excretion of the right leg has smwt decreased, edema throughout too
per os taken some drinks and a little bit of gruel in the afternoon

…

DISCHARGE DOCUMENT
REASON FOR THE ICU ADMISSION AND ANAMNESIS
Verbatim copy of the admission document, except *O room* 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 tubes, sufficient extubation with the 40/50% ventimask. 24.3 with 28% VM oxygen 9.5 and CO2 5.6. 23.3 in the THX- scan increased pleural fluid and mild incompensation scan.Continued hevy minussing (DRUGNAME 10mg x 6 iv.) 24.3 the amount of pleural fluid gone down. DRUGNAME stopped at11, will be given when needed according to the response.
HAEMODYNAMICS: Pulse tachycardic,extrasystoles, flimmer. Adission RR low. DRUGNAMES infusions of a large dosage. Filling. However, CI quite low. IABP 1:1. 20.3. started DRUGNAME, when the rhythm did not convert with electricity. 21.3. DRUGNAME stopped As a new finding, left branch dissociation, which improved 20.3 C.I 1.5 --> started DRUGNAME infusion. Currently RR even too high. Pulse: FA, tachycardic.
CONSCIOUSNESS AND MOOD: ICU started DRUGNAME infusion, got DRUGNAME boluses too. DRUGNAME 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 -Ukg performed by H. Suominen: A clear liquid diaper around the heart , good contraction.*
¤ *20.3 esophagus- Ukg / K. Haverinen: Septum is faint, contracts mod. Mitral- and aortic valve ok. No explanation for LHS branch dissociations been found.*
INFECTION SITUATION: 23.3 CRP 32
EXCRETION:
Profuse diuresis, because DRUGNAME 10 mg iv every 4h.26.3 Diuresis at 6 -->: 12 ->1090 ml
Motion: -
Drains: After o profuse drainbleeding, Hb low, needed a lot of red cs.In the esophagus-ukg observed liquid diaper around the heart and impending tamponation. Obsed the situation until next 8 a.m., when the cnclsn was resternomy, where found 2 points 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 DRUGNAME + PRODUCTNAME. Skin sensitive all over; nicks caused by tapes etc.
PAIN MANAGEMENT: DRUGNAME, which makes very tired.
SPECIAL CARE: 19.-23.3. IABP
RELATIVES: Two sons been in contacts, 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 -> insulin DRUGNAME-infusion. Insulin cut out for the transit.
- Potassium high and changed PL-K to Na 0.3 at 12.

Figure 2.2: *Style-preserving illustration of Finnish intensive care nursing narratives translated to English*

The analysis addresses nursing narratives, because they cover the whole ICU in-patient period and are recorded at all times, during every shift. Nursing documentation is based on the process of gathering information from the patient, setting goals for care, documenting nursing interventions and evaluating the delivered care. All clinicians of multi-professional ICU team use nursing narratives in decision making.

The ICU nursing narratives can be grouped with respect to the time of writing into *admission documents*, *daily nursing notes*, and *discharge documents* (Figure 2.2). An admission document is a compact patient description at the time of ICU admission. Daily nursing notes are written during the actual ICU stay and stored into patient- and nursing shift-specific documents. They are mostly used for information exchange among the clinicians in the unit. In this thesis, the file containing all daily nursing notes of the patient is called a *daily nursing note document*. Discharge documents aim at transferring summarized information from the ICU to other wards to assure the continuity of care.

2.3 Patient record data

To specify clinical needs for HLT, we analyse narratives written in a Finnish ICU for adults (Figures 2.3 and 2.4). The ICU is located in a university-affiliated hospital and is a 24-bed mixed medical closed unit accommodating approximately 2,000 admissions per year (1,815 in 2006). In 2006, the average length of stay (LOS) was 3.3 days with a standard deviation of 4.9 days. The normal, planned strength of the nursing personnel is twenty on the morning shifts, eighteen on the evening shifts and twelve on the night shifts. In comparison, the unit has at least two ICU physicians constantly on duty. The ICU has been using electronic patient records since 2002.

Our data includes nursing narratives from 516 patient records between 2005 January 1 and 2006 August 1. They have been collected with proper permissions retrospectively and de-identified (see Section 2.1). Due to the development of documentation practices, the most recent patients were selected; collecting the data started on 2006 August 1. Our inclusion criterion is LOS of at least five days, because fluent information flow is likely to be challenged with protracted in-patient periods. The number of records fulfilling this criterion is 516 out of the total number of $2,789$ patients that have been taken care of in the ICU during the time period.

Photo: Heljä Lundgrén-Laine

Figure 2.3: *Finnish intensive care*

Photo: Heljä Lundgrén-Laine

Figure 2.4: *Multi-professional and multilingual intensive care (note a Finnish-Swedish-Finnish dictionary in the bottom)*

The writing style in the nursing narratives is telegraphic and highly domain-specific with a substantial amount of unit-specific terminology, headings and other documentation practices (Figure 2.2). Admission and discharge documents are guided by default headings in order to direct the use of a specific structure. However, the personnel can remove and modify the headings, as well as use other headings. There are no default headings for the daily nursing notes.

In addition to nursing narratives, the data include the following structured or numerical demographics about the patients: Structured data are

- admission type (449 emergency, 67 elective),
- main ICU diagnostic group with respect to the International Classification of Diseases, 10th Revision (ICD-10, 115 individual groups are present in the data),
- sex (183 female, 333 male),
- ICU outcome (398 recovering, 21 in the middle of care, 29 no care outcomes, 57 dead, 11 missing value), and
- ward to which the patient was discharged (155 medical ward, 242 surgical ward, 56 other hospital, 56 dead, 1 missing value).

Numerical data are

- age (minimum 6, maximum 88, mean 59, standard deviation 16),
- LOS in days (minimum 5.0, maximum 58, mean 11, standard deviation 7.3), and
- average nursing intensity scores (referred to as score, i.e., evaluations of care needs of the patient, varies from one (i.e., low need for care) to five (i.e., a high need for care), minimum 2.2, maximum 4.6, mean 3.2, standard deviation 0.37).

2.4 Analysis aspects and methods

The first analysis aspect is the amount of narratives per patient and demographics that contribute to the large amount, and thus challenge the information flow. For instance, the admission type could be an explanatory factor because of considerable differences between the groups of elective patients being prepared for a planned operation and emergency patients.

The amount of text per patient is studied through the minimum, maximum, mean and standard deviation of the number of words in admission,

daily nursing note, and discharge documents. For daily nursing note documents, not only the total amount but also seven other number-of-words variables are considered: (1) the average amount written per day; (2–5) the cumulative amount written during the first one, two, three and four days of stay; and (6, 7) for patients with long enough LOS, the amount written during the first week and two weeks of the stay.

Demographics that contribute to the large amount are specified statistically. We divide the 516 patients into five approximately equally sized groups according to the number of words in the daily nursing note document. Thus, as an example, patients belonging to group 1 have the shortest daily nursing note documents and those in the following group have the next smallest amount of documentation. In addition, because there are so many diagnostic groups present and consequently only few patients per group, we merge the groups into two granularities using the ICD-10 tree and features of diseases as guidelines. The first granularity contains nine disjoint groups:

1. *Infections* (36 patients),
2. *Tumours* (23 patients),
3. *Endocrinology, nutrition and metabolism* (8 patients),
4. *Diseases of muscle and nervous system* (29 patients),
5. *Cardiovascular diseases and problems in conduction system* (217 patients),
6. *Lung diseases* (64 patients),
7. *Diseases of the abdominal cavity organs* (42 patients),
8. *Unclassified symptoms or abnormal clinical and laboratory findings* (11 patients), and
9. *Traumas, intoxications and extrinsic factors* (86 patients).

The second granularity has four disjoint groups:

1. *Cardiovascular diseases and diseases of muscle and nervous system* (i.e., old 4 and 5, 246 patients),
2. *Lung diseases and infections* (i.e., old 1 and 6, 100 patients),
3. *Diseases of the abdominal cavity organs and metabolism problems* (i.e., old 3 and 7, 50 patients), and
4. *Other* (i.e., old 2, 8 and 9, 12 patients).

Logistic regression is performed separately for the two granularities of diagnostic groups but all other demographics are the same in both analyses, and the procedure is repeated for the seven other number-of-words variables. Then, the eight number-of-words variables with the respective four-level diagnostic groupings are compared using the Kruskal-Wallis test due to the skewed distributions. In further pairwise comparisons with the Mann-Whitney U, the Holm variation of the Bonferroni correction is applied. All statistical analyses are performed with SAS 9.1. For information regarding the statistical tests, see, for example, (Norman and Streiner, 2000, pp. 71, 139-144, 225, 226).

The second analysis aspect is the content. As structuring eases efficient accessibility of narratives, headings are studied. In these analyses, content analysis is used. This method is defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding, and it also allows for monitoring shifts and changes in documentation style and content (Miles and Huberman, 1994; Stemler, 2001). Various heading spellings with the same meaning are combined into one heading. A similar methodology of comparing headings with their content is used in (Hyun and Bakken, 2006) with English nursing narratives from the USA.

Manual and semi-automated techniques are used in parallel. For the completely manual phase, three admission, daily nursing note, and discharge documents are chosen so that they reflect the variability of the narratives. The documents include

1. the admission, daily nursing note, and discharge document of the patient on whom the largest amount of narratives is stored (the admission document and daily nursing note document of this patient were the largest), as the problems related to the utilization of stored information are likely to be more evident when the amount of narratives is large,

2. the shortest admission, daily nursing note, and discharge document to observe the contrast between long and short narratives, and

3. the admission, daily nursing note, and discharge document corresponding to the patient whose daily nursing note document is closest to the average size of this document type.

The headings in all daily nursing note documents are studied semi-automatically, because default headings are not used to guide this part of the documentation. For this purpose a heuristic based on the assumption that a heading is separated from the text by a colon and a space character is applied. The heuristic has 99 percent precision and 90 percent recall (see

Section 3.1.3) when applied to the three daily nursing note documents considered in the manual phase.

2.5 Problems and challenges in the information flow

2.5.1 Amount

An overwhelming amount of nursing narratives (Table 2.1) challenges the fluency of the information flow: The largest admission, daily nursing note, and discharge document had about seven, 48, and four pages (A4 paper, 3 cm margins on every side, Times New Roman font, font size 12, single line spacing, i.e., approximately 270 words per page), respectively. On average, 3,000 words (c. 11 pages) of nursing narratives were written about a patient during the entire ICU in-patient period the daily amount being about 200 words. The average size of the admission document was about 250 words and the respective numbers for daily nursing note and discharge documents were 2,100 and 400. In conclusion, the volume of text is challenging for efficient accessibility and hence, there is a risk of not using all gathered information in patient care. In addition, summarizing the content or creating overviews and trends about a certain aspect must be laborious and difficult.

However, the amount of text per patient varied substantially. As an example, the five shortest daily nursing note documents contained 480, 590, 600, 620, and 640 words each, whereas the five longest daily nursing note documents had 8,600; 9,600; 12,000; 12,000; and 13,000 words each. This creates motivation for specifying the demographics that contribute to the large text amount.

Logistic regressions (Table 2.2) gave evidence of a high nursing intensity contributing to the large amount of daily nursing notes. The interpretation is that when the patient has a high need for nursing care, there are also many things to document. Contrary, more daily nursing notes were written about the elective patients than about the emergency patients especially in the beginning of the ICU stay; the admission of elective patients to the ICU is arranged beforehand and the medical status of these patients is carefully checked in advance, and thus, it would have been expected that emergency patients have more things to be documented. This finding may illustrate a potential problem in the information flow: It seems that in the beginning of the ICU period, emergency patients require such intensive nursing care that the personnel has no time for writing nursing narratives. As expected, there was a significant positive relationship between the total amount of daily nursing notes and LOS.

Table 2.1: *Descriptive statistics for the amount (in words) of nursing narratives*
Reprint from Suominen et al. (2009a)

Document type	***n***	**Mean**	**Standard deviation**	**Minimum**	**Maximum**
Admission document	348	250	240	8	2,100
Daily nursing notes: 1st day	516	270	110	37	720
Daily nursing notes: 1st 2 days	516	480	170	130	1,290
Daily nursing notes: 1st 3 days	516	690	220	240	1,700
Daily nursing notes: 1st 4 days	516	880	260	290	2,000
Daily nursing notes: 1st week	382	1,500	360	660	2,800
Daily nursing notes: 1st 2 weeks	99	2,800	590	1,600	4,600
Daily nursing notes: Total	516	2,100	1,500	480	13,000
Daily nursing notes: Daily average	516	190	44	89	380
Discharge document	514	400	140	140	1,000

Table 2.2: *Results of logistic regressions aiming to identify the demographics that increase the amount of daily nursing notes*

Reprint from Suominen et al. (2009a)

Lists only the effects with the 95% Wald confidence limit strictly below one or strictly above one.

Groups refers to the diagnostic group granularity, score to the average nursing intensity score, and em. and el. to the emergency and elective admission types.

Amount	**Groups**	**Effect**	**Odds ratio**	**95 % Wald confidence limit**
1st day	4	score	3.041	(1.904–4.857)
1st day	4	em. vs. el.	0.579	(0.353–0.950)
1st day	9	score	2.688	(1.699–4.330)
1st day	9	group 2 vs. 9	0.380	(0.157–0.919)
1st 2 days	4	score	2.645	(1.659–4.216)
1st 2 days	9	score	2.401	(1.492–3.864)
1st 3 days	4	score	6.364	(3.890–10.412)
1st 3 days	9	score	5.659	(3.431–9.331)
1st 3 days	9	em. vs. el.	0.479	(0.344–0.976)
1st 3 days	9	group 2 vs. 9	0.313	(0.128–0.765)
1st 3 days	9	group 6 vs. 9	0.441	(0.228–0.852)
1st 4 days	4	score	6.600	(4.023–10.828)
1st 4 days	4	em. vs. el.	0.563	(0.341–0.929)
1st 4 days	9	score	6.300	(3.799–10.449)
1st 4 days	9	em. vs. el.	0.535	(0.554–0.904)
1st week	4	score	7.982	(4.282–14.877)
1st week	9	score	8.733	(4.757–16.033)
1st week	9	group 6 vs. 9	0.340	(0.162–0.752)
1st 2 weeks	4	score	35.046	(7.774–157.989)
1st 2 weeks	9	score	44.819	(8.922–225.149)
1st 2 weeks	9	group 3 vs. 9	0.013	(< 0.001–0.445)
1st 2 weeks	9	group 6 vs. 9	0.073	(0.013–0.412)
Total	4	score	5.007	(2.826–8.872)
Total	4	LOS	2.880	(2.540–3.260)
Total	9	score	4.699	(2.621–8.425)
Total	9	LOS	2.889	(2.551–3.270)
Total	9	group 2 vs. 9	0.292	(0.100–0.854)
Total	9	group 6 vs. 9	0.380	(0.170–0.851)
Daily average	4	score	7.658	(4.642–12.633)
Daily average	9	score	7.070	(4.247–11.768)
Daily average	9	group 2 vs. 9	0.310	(0.127–0.757)
Daily average	9	group 6 vs. 9	0.349	(0.180–0.679)

According to Kruskal-Wallis tests comparing the amount of daily nursing notes and the rougher diagnostic grouping, the largest amount of daily nursing notes was written about patients belonging to group *4. Other*. The groups with the second and third largest amount were *1. Cardiovascular diseases and diseases of muscle and nervous system* and *3. Diseases of the abdominal cavity organs and metabolism problems*, respectively. The smallest amount of daily nursing notes was written about patients of group *2. Lung diseases and infections*. The order of the four groups was always the same when the relationship between the amount of daily nursing notes and diagnostic groups was statistically significant ($p < 0.05$, all comparisons except the total amount). The pairwise comparisons verified that the amount of daily nursing notes written about group 2 was statistically significantly smaller than that about groups 4 and 1. The difference between groups 3 and 4 was statistically significant in the case of the amount of daily nursing notes written during the first week of care.

2.5.2 Content in admission documents

The default headings in the admission documents were

1. *Previous/other diseases and anamnesis*,
2. *Reason for admission*,
3. *Breathing*,
4. *Hemodynamics*,
5. *Diuresis*,
6. *Excretion*,
7. *Consciousness and mood*,
8. *Nutrition*,
9. *Pain management*,
10. *Skin and wound care*,
11. *Medical treatments*,
12. *Infections*,
13. *Special treatments*,
14. *Rehabilitation*,
15. *Belongings*,

16. *Relatives*,

17. *Other*.

In the longest admission document (about 2,100 words), precisely these seventeen headings were used. Under each heading, several topics were discussed, and the same topic was often discussed under multiple headings as the most common topic in this document, patient's medical problems, illustrates. However, when a particular topic was discussed under multiple headings, the nurse's viewpoint differed. For example, nurses discussed intracranial pressure (ICP) under several headings. But under *Hemodynamics* the effect of the heart rhythm on ICP was discussed, under *Consciousness and mood* the optimum levels of ICP in relation to consciousness were described, and under *Pain management* ICP values were used when estimating analgesic actions. These observations were consistent with the analysis of the admission document (about 420 words) of the patient with the average size daily nursing note document. The only content in the shortest admission document (8 words) was the patient ID, the default heading *Previous/other diseases and anamnesis* and under it the phrase *Previously been healthy.* In conclusion, this topical scattering challenged the fluency of the information flow; to be able to create a general impression of ICP, for instance, requires collecting all the related pieces and perspectives and performing this manually is time-consuming.

2.5.3 Content and information flow in daily nursing notes

On the basis of the semi-automatic content analysis of 516 daily nursing note documents, headings were used to structure daily nursing notes on only half of approximately 18,400 shifts in total. However, the headings seemed to be quite well-established even though default headings were not automatically given; the most commonly used headings were

1. *Hemodynamics* ($n \approx 7,800$),

2. *Consciousness* ($n \approx 6,900$),

3. *Relatives* ($n \approx 5,700$),

4. *Diuresis* ($n \approx 5,400$),

5. *Breathing* ($n \approx 4,500$),

6. *Oxygenation* ($n \approx 3,600$), and

7. *Other* ($n \approx 3,200$).

Other headings were used fewer than 600 times per heading. Of these, only four were mentioned more than 100 times:

8. *Excretion* ($n \approx 590$),
9. *Hemodialysis* ($n \approx 370$),
10. *Pulse* ($n \approx 160$), and
11. *Skin* ($n \approx 160$).

However, as each heading had a considerable number of synonyms and spellings, topical search from daily nursing notes seems even more problematic than from admission documents.

To illustrate the substantial amount of different spellings in the vocabulary, originally at least 140 various spellings of the word *hemodynamiikka* [*hemodynamics*] were present in daily nursing notes, when all different capitalizations, spellings, and inflections were calculated. The nominative *hemodynamiikka* was spelled in over thirty different ways: *hemo*, *hemod*, *hemodnamiikka*, *hemodunamiikka*, *hemoduynamiikka*, *hemodyamiikka*, *hemodyanamiikka*, *hemodyanmiikka*, *hemodybamiikka*, *hemodyhamiikka*, *hemodyknamiikka*, *hemodymaniikka*, *hemodymiikka*, *hemodyn*, *hemodynaamiikka*, *hemodynaamikka*, *hemodynakiikka*, *hemodynam*, *hemodynamiiakka*, *hemodynamiiikka*, *hemodynamiika*, *hemodynamiikk*, *hemodynamiikka*, *hemodynamiikki*, *hemodynamkiikka*, *hemodynammiika*, *hemodynaniikka*, *hemodynmiikka*, *hemodynnamiikka*, *henonynamiikka*. As another example, the medicine *noradrenaliini* [*noradrenaline*] had over 350 various spellings in the daily nursing notes and over sixty variants of the nominative: *nonadrenalin*, *nonadrenalina*, *nor*, *norad*, *noradenaliini*, *noradenalina*, *noraderenalina*, *noraderenalini*, *noradernalina*, *nor-adr*, *nor.adr*, *noradr*, *noradranalian*, *noradranlina*, *noradreanliini*, *noradreanlina*, *noradreliini*, *noradrelin*, *noradren*, *noradrenal*, *noradrenalaina*, *noradrenali*, *noradrenalia*, *noradrenalian*, *noradrenaliia*, *noradrenaliiina*, *noradrenaliimi*, *noradrenaliin*, *noradrenaliina*, *noradrenaliini*, *noradrenaliinna*, *noradrenalin*, *noradrenaline*, *noradrenalini*, *noradrenallina*, *noradrenallini*, *noradrenanil*, *noradrenanilin*, *noradrenenalina*, *noradrenliini*, *noradrenlin*, *noradrenlina*, *noradrenlini*, *noradrnaliini*, *noradrnalina*, *noradrrenalin*, *noradrtenalin*, *norandrenalina*, *norarадrenalina*, *norarenaliini*, *norarenalina*, *noratrenaliini*, *nordarenaliini*, *nordrenaliini*, *nordrenalin*, *nordrenalina*, *nordrenanili*, *noreadr*, *norradenalina*, *norradrenaliini*, *norradrenalin*, *norradrenalini*, *norri*, *norrradrenalin*.

Table 2.3: *Headings, their frequencies and words accumulation in the longest daily nursing note document*

Reprint from Suominen et al. (2009a)

Heading	**1st 1/4**	**2nd 1/4**	**3rd 1/4**	**4th 1/4**
Hemodynamics	23	19	17	9
Relatives	21	14	15	7
Consciousness	11	9	15	6
Diuresis	18	12	7	5
Oxygenation	13	13	9	5
Breathing	9	7	10	
Other	8	5	7	
Neurology	5	5	2	
Excretion	3	3	3	
Neurological status	3	4		
Consciousness and ICP		2		
Consciousness and mood	1		1	
Heart and blood circulation	2			
Infections			2	
Basic care		1		
Brain pressure	1			
EEG	1			
Freezing therapy	1			
Head wound	1			
Pain management	1			
Rehabilitation and mood			1	
Sedation	1			
Situation of the head	1			
Skin and wound care	1			
Special treatments	1			19
Treatments	1			99
No headings	6	9	10	72
Column sum	133	103	11	2,400
Words	4,300	3,900	10	2,800

The result of the manual content analysis of the daily nursing note document (about 13,000 words) corresponding to the patient with the largest amount of narratives verified the results of the semi-automated phase; the document had several shifts where no headings were used and the most often used headings were 1. *Hemodynamics*, 2. *Relatives*, 3. *Diuresis*, 4. *Consciousness*, 5. *Oxygenation*, and 6. *Breathing* (Table 2.3). The systematic use of the same headings in time can be interpreted as an evidence of a fluent information flow from a shift to another even though this analysis did not address information flow within daily nursing notes in detail. The number of headings nurses used decreased during in-patient period, and in the last quarter, only the seven most common headings were used[1].

In the about-13,000-word daily nursing note document, the topics nurses discussed did not change whether headings were used to structure the text or not, but when not using headings, nurses wrote more verbally, and grammatically correctly with a focus on basic care. The most common topics were hemodynamical problems, skin condition, patient's communication, and the personnel's conversations with the patient's relatives. During the last quarter, the emphasis was on the patient's vital signs and the supportive work concerning the close relatives of the patient. The amount of notes decreased over time as the text seemed to get stylistically more and more telegraphic (e.g., *Blood pressure and pulse unchanged*).

The problems related to topical scattering were evident. First, altogether 407 individual headings were used the 13,000-word daily nursing note document. Second, nurses used headings which covered similar issues but were not exactly synonymous, such as the pairs *Oxygenation–Breathing* and *Hemodynamics–Blood circulation*. Third, as in the admission documents, similar topics were discussed under several headings. For instance, ICP was again discussed under multiple headings but from various perspectives. Fourth, nurses often documented matters that can be seen as unrelated to the heading. As an example, the topics under the heading consciousness varied from medication to rehabilitation.

The daily nursing note document of average length (about 2,100 words) had 92 headings in total. The headings most often used were 1. *Hemodynamics*, 2. *Relatives*, 3. *Consciousness*, 4. *Breathing*, 5. *Diuresis*, and 6. *Other*. The other headings used were *Sedation and consciousness*, *Skin lesion*, *Drains*, *Oxygenation*, *Pulse*, and *Excretion*. The stronger homogeneity in the headings and topics of this daily nursing note document was probably due to the smaller number of writers.

In the shortest daily nursing note document (about 480 words), the topics nurses discussed did not change whether headings were used, and the head-

[1]The document was divided into shift-wise quarters (i.e., shifts 1–30, 31–60, 61–90, 91–120).

ings were not used on one third of the shifts. The most common headings were consistent with the headings of the longest document: 1. *Hemodynamics*, 2. *Relatives*, 3. *Consciousness*, 4. *Breathing*, and 5. *Diuresis*. Parallel heading pairs were also found.

2.5.4 Content and information flow in discharge

The default headings in discharge documents were

1. *Reason for admission and anamnesis*,
2. *Breathing*,
3. *Hemodynamics*,
4. *Consciousness and mood*,
5. *Medical treatments*,
6. *Infections*,
7. *Nutrition*,
8. *Excretion*
    i. *Diuresis after 6 a.m.*,
    ii. *Defecation*,
    iii. *Drains*,
9. *Skin and wound care*,
10. *Pain management*,
11. *Rehabilitation*,
12. *Special treatments*,
13. *Relatives*, and
14. *Belongings*.

In the discharge document (about 710 words) of the patient with the largest amount of narratives, all default headings were used. In contrast to admission and daily nursing note documents, the same issue was mostly documented under only one heading (e.g., ICP under *Consciousness and mood*, oxygenation under *Breathing*) and under each heading only topics closely related to the heading were discussed. The focus was on medical and vital problems. The structure and content of the document was the same as in the

admission document of this patient, and the headings *Breathing*, *Consciousness and mood*, and *Rehabilitation* contained the largest amount of notes. Even though it is essential from the information flow perspective that the content of the admission document reaches the discharge phase, this document more adequately illustrated a problem in the flow: the phrases were copied almost word for word from the admission document and the content of the daily nursing note document was extremely poorly discussed. This is probably due to the substantially large amount and incoherent structure of daily nursing notes.

In the discharge document (about 450 words) of the patient with the average size daily nursing note document, the text was mainly directly copied from the admission document. A few notes concerning the patients breathing, consciousness, laboratory tests, and belongings were added. Texts from daily nursing notes were not directly utilized.

The shortest discharge document (about 150 words) had all the default headings but almost half of them included no notes. The writing style was reminiscent of a checklist as, for example, some of the vital functions were described only with the word OK. The document contained the patient's state at the time of discharge, but no other utilization of daily nursing notes was observable.

2.5.5 Summary

The information flow in ICU nursing narratives is currently fragmented: The first challenge is the overwhelming amount of text. This is likely to be particularly evident, when the in-patient period is protracted, the patient has a high need for nursing care, elective admission type, or belongs to the diagnostic groups of *Tumours*, *Unclassified symptoms or abnormal clinical and laboratory findings*, or *Traumas, intoxications and extrinsic factors*. Second, it seems that clinical needs for HLT vary with patient demographics. For example, support is needed in particular for documenting the care of emergency patients. Third, the headings in daily nursing notes are lacking a systematic naming, and notes in admission documents and daily nursing notes are topically scattered. This complicates both the information access and intelligibility. Finally, there is clear evidence of information flow problems and support is needed to all phases: from admission documents to daily nursing notes, from admission documents to discharge documents, within daily nursing notes, and from daily nursing notes to discharge documents (Figure 2.5).

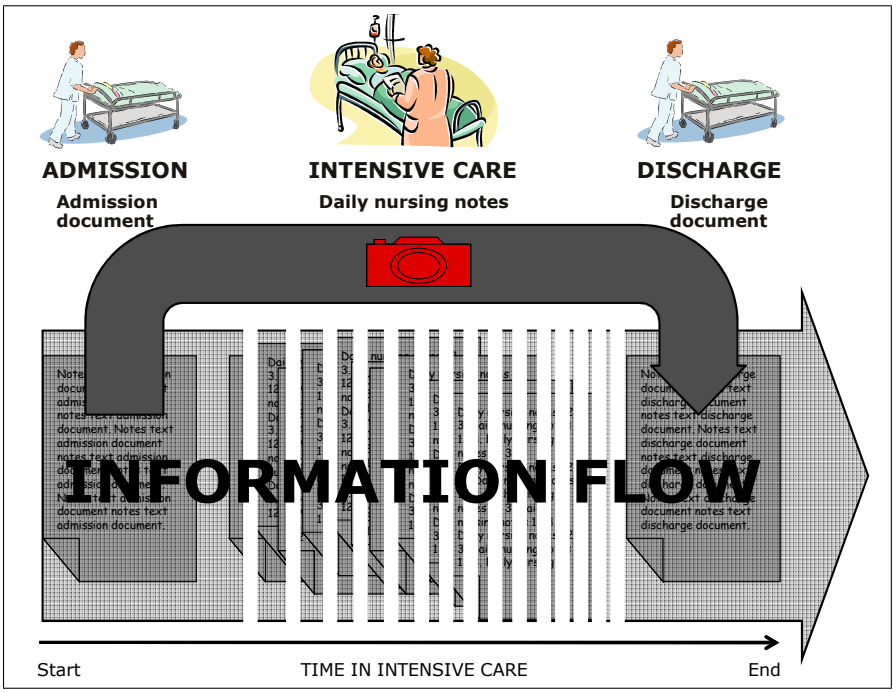


Figure 2.5: *Current problems of the information flow: copying and pasting admission documents to discharge documents and emphasized fragments in time in using daily nursing notes*

2.6 Road map for human language technology development to support the information flow

Next, I introduce a road map (Figure 2.6) for developing HLT to support the fragmented information flow. Fragments in the information flow are related to using previously gathered texts in the later phases of the patient care process. This problem can be generalized from the documents of a given patient to also using text from previous patients. Hence, the road map includes the data resource of the *evidence-based practice repository* in addition to the *patient's admission document*, *daily nursing notes*, and *discharge document*. The repository contains both scientific research evidence and the practice-based evidence represented in the previous patient records. It supports information flow from the previous patient cases to the care of the new patient and enhances capabilities to deliver evidence-based care. The HLT components of the road map arise from the previously depicted problems and challenges. The components include patient profile building, supporting the writing, attention-focusing, summarizing, and proof-reading.

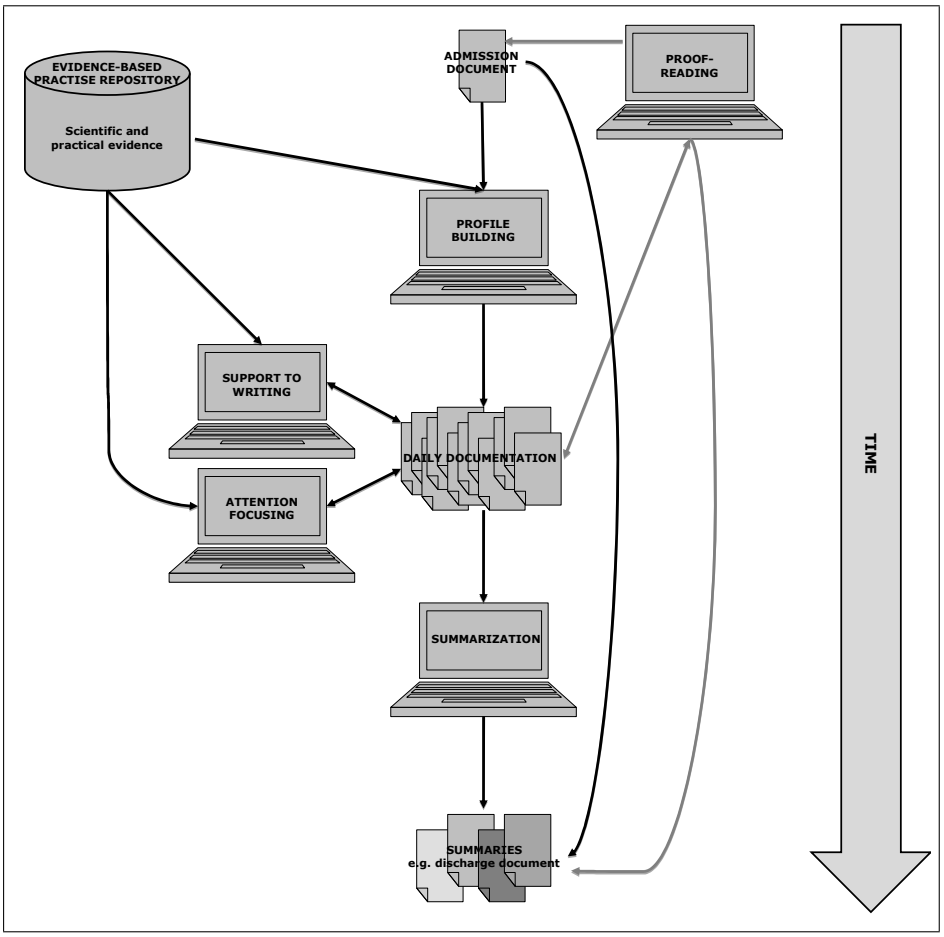


Figure 2.6: *Road map for developing human language technology*

At the admission phase, the *profile building component* creates an individual patient profile by comparing the admission document with the evidence-based practice repository. This component arises from different patient demographics affecting the content and amount of narratives. For example, the profiles of patients with different diagnoses may differ. In the later phases, this individual profile can be used, for instance, to determine a free-text documentation structure addressing the topics that are particularly relevant for the profile. A more advanced example is to automatically monitor the notes written during the in-patient period, and to generate reminders about, for instance, topics that are relevant for the patient profile but do not seem to be addressed yet.

Next, the phase with daily nursing notes is addressed with the *support the writing component*. As discussed above, the patient profile and evidence-based practice repository can be used together to direct the writing of daily nursing notes suitable for the particular patient characteristics. The HLT component is first used for generating a suitable document structure by comparing the patient profile of the new patient with the knowledge about previous patients. The structure contains the headings that are generally used in daily nursing notes of similar patients. This standardizes the headings, but the structure should be freely modifiable and its use optional. The HLT component is then used for providing real-time documentation feedback. The component analyses the correspondence of daily nursing notes with the patient profile and the evidence-based practice repository, and based on the analyses, provides feedback. For example, if management of pain is a common topic in the documents of previous similar patients, the component may remind the user to document this topic. In conclusion, the component that supports writing daily nursing notes promotes individualized care.

HLT can also be used for attention focusing during the in-patient period. The *attention-focusing component* supports the use of previously gathered narratives by analyzing the content of daily nursing notes and generating statistics, trends and visualizations of both numerical and narrative data. The component also compares daily nursing notes of the new patient with the knowledge about previous patients, and generates alerts about potential threats. For example, an alarm about the risk of pressure ulcer could be launched if notes about redness of skin have been recorded during the previous shift. In addition, the component supports searching notes relevant to a certain topic. The purpose of the component is to provide tools to outline and predict the patients trajectory over time. In summary, the attention focusing component supports direct clinical decision making and evidence-based practice.

Finally, at the discharge phase and also already at earlier phases of the ICU in-patient period, HLT can be used for summarization. The *summarizing component* contains an interactive tool that analyses the content of the admission document and daily nursing notes. The component first helps in specifying the summarization topics based on the patient profile and evidence-based practice repository. Also the target audience of the summary can be taken into account, as the information needs vary from a ward to ward (e.g., ICU, medical ward, traumatology unit, or pathology ward) and from a profession to profession (e.g., general practitioner, heart specialist, nurse, physiotherapist, or the patient). Fully automated summarization being an extremely difficult task, the component provides functions to search information relevant to the user-specified topics and highlight the results in their actual context. When the technology matures, steps towards fully automated summarization can be taken.

To make these advanced HLT components that support information access work, *proof-reading components* are needed. The knowledge stated in the narratives is not readily available for further automated analysis due to the ambiguous nature of human language, where the meaning of a sentence depends on its context and, on the other hand, a single meaning can be expressed in a number of equivalent ways. A crucial step in resolving this ambiguity is to find the correct syntactic structure of each sentence so as to explicitly capture word and phrase relationships that are not evident from the linear order of words. Parsing, the process of syntactic analysis of natural language sentences, is thus a critical prerequisite for the majority of advanced HLT components. It allows a deeper analysis of the text than would be otherwise possible. In addition, performance of all HLT components can be improved by developing spelling correction programs. Further performance improvements can be gained by reducing text sparseness by using and constructing domain-adopted terminologies through HLT as well as increasing the intelligence of the proof-reading components in guiding the author (Figure 2.7).

In conclusion, the proof-reading components are beneficial for other HLT components but more importantly they contribute to text intelligibility for health care professionals and the patients themselves. Intensive care is a multi-professional domain, and thus several different profession groups write and read narratives. These different profession groups may have difficulties in understanding each others' language. Further difficulties occur, when patients are discharged to other hospital wards and the non-ICU personnel read discharge documents. Finally, patients often have problems in understanding their own patient narratives (see Figure 2.2) even though the legislation requires general intelligibility and clarity of patient records (Decree 99/2001 of the Ministry of Social and Health, Finland). Here language can be understood as a profession or ward-specific jargon (e.g., ICU or medical ward nursing Finnish in Turku), or dialect (e.g., Finnish in Southwest Finland or Eastern Finland) or as a language in its general meaning (e.g., Finnish or English). With HLT, the language in electronic health records can be brought nearer to the language of the target audience, and further support can be given by offering term definitions from terminologies and references for related care guidelines, studies, and other similar references. This increased intelligibility supports not only health care professionals in decision making but also cognitively empowers individuals to actively make decisions concerning their health.

When applied to the evidence-based practice repository, the above discussed HLT components can be used to generate statistics, trends, visualizations, summaries and other overviews about patient groups. This capability would benefit not only direct care but also health administration, research, and education.

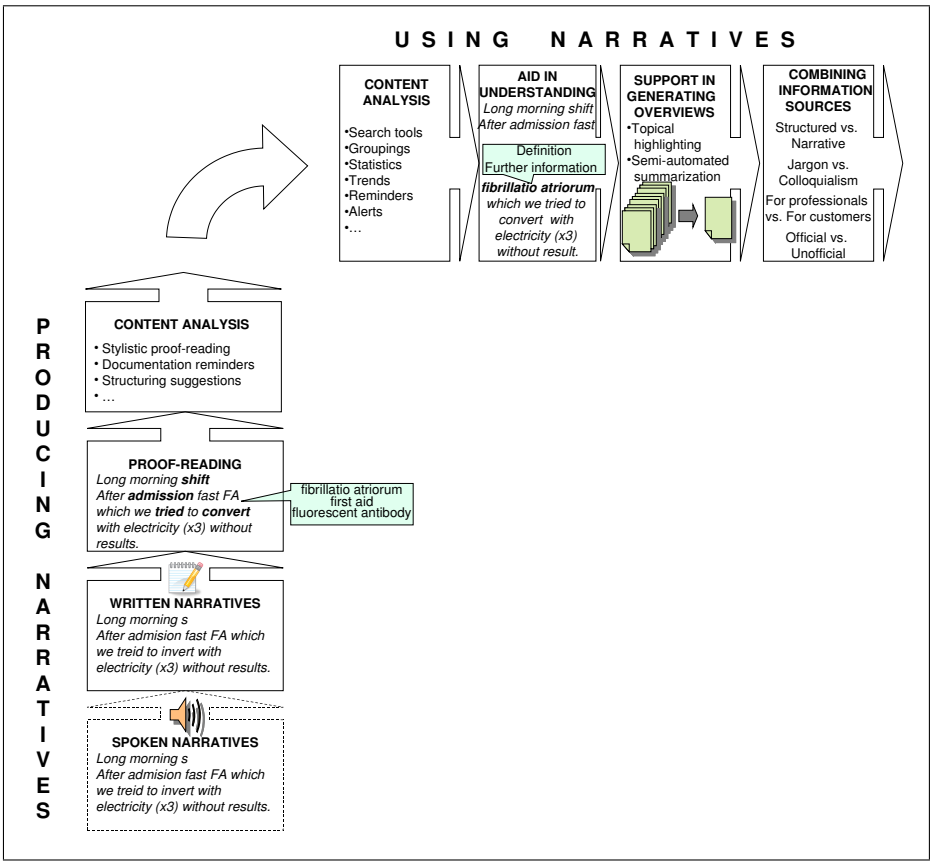


Figure 2.7: *Summary of the support for producing and using narratives*

In summary, the HLT components offer many functions for producing and using narratives (Figure 2.7). This supports both spoken and written health communication, and personalized (i.e., individual profiles), predictive (i.e., reminders and alarms), preventative (i.e., using existing records and gathering new evidence), participatory (i.e., individuals and multiprofessional health care provider team) 4P health care.

2.7 Comparison with the previous work

The road map is consistent with studies related to clinical documentation and the use of HLT. Let us consider the clinical motivation and data resources first. By analyzing altogether 70 patient records from Finnish surgical, neurological, maternity, and childrens wards, Kärkkäinen and Eriksson (2003) show that nursing narratives are detailed descriptions of the

nursing care process and evaluating their content is exhausting and time-consuming. Further, based on the analysis of 35 Thai patient records from a medical-surgical ward, crucial information is lacking from nursing narratives in terms of topics and sufficiency of notes; narratives have inconsistencies, irrelevant notes, and time-line gaps; and unsuitable default headings lead to topical scattering, duplicated information, and increased documentation time (Cheevakasemsook et al., 2006). Moreover, Hellesø (2006) analyses narrative admission documents, care plans and discharge documents of 66 patients from Norwegian medicine and cardiopulmonary units and concludes that broad default headings help nurses improve the completeness, structure, and content of discharge documents, but the headings should differentiate between various patient demographics. Finally, according to the comparison of headings and their content in 43 different types of US nursing narratives from a US hospital, the naming and content of headings is inconsistent, and the same heading is used in various documents (e.g., Nursing Adult Admission History and Post Anesthesia Care Unit Discharge Criteria) (Hyun and Bakken, 2006). In summary, admission, daily nursing note, and discharge documents contain valuable information, but its flow within and between ICU admission and discharge needs support. For example, Weaver et al. (2005) is a study related to the evidence-based practice repository. It presents a strategy of using information technology as the underlying tool to capture empirical evidence from electronic health records.

Let us now consider the HLT components. Proof-reading and terminology building are addressed, for instance, in Ruch et al. (2003), Zhou et al. (2006), and Laippala et al. (2009). Viewpoints connected to profile building are studied by identifying patients with a given diagnosis, for instance, in Pakhomov et al. (2006), Hripcsak et al. (2007), and Paper V. Paper III, Jancsary and Matiasek (2008), and Paper IV are examples of studies with a focus on documentation structure and topical search, and Meystre and Haug (2006b), Zeng et al. (2006), and Pentz et al. (2007) develop methods for extracting information from electronic health records. These studies are also relevant for the summarization component. For example Gundlapalli et al. (2007) addresses generating alerts. See, for example, Friedman and Hripcsak (1999), Zweigenbaum et al. (2007), and Meystre et al. (2008) for a more thorough review of related work. In conclusion, collaboration and a road map to follow are needed to integrate these numerous distinct tools to a compressive HLT solution.

When comparing the entire road map with previous work, similar ideas are described in Goodwin et al. (2003) and Heldt et al. (2006). Goodwin et al. (2003) provides a model for building nursing knowledge with data mining methods, but it does not explicitly include HLT and it is from the perspective of the clinical environment on a more abstract level. The road map in Heldt et al. (2006) is technically more detailed than ours: It contains

the major modules of an advanced patient monitoring system aiming to improve the efficiency, accuracy and timeliness of clinical decision making in the ICU. Although clinical notes are explicitly considered, it does not identify HLT interfaces.

Chapter 3

Machine Learning Applications

This chapter presents three practical cases of applying machine learning to support the information flow: The first two cases, that is, (1) topic labeling and relevance ranking and (2) topic segmentation and labeling, ease information search and the third one is for diagnosis coding. The case of topic labeling and relevance ranking has two supervised machine learning applications: binary classification and regression. The case of topic segmentation and labeling includes both a supervised and unsupervised multi-class classification application. The case of diagnosis coding discusses a supervised multi-label classification application.

3.1 Topic labeling and relevance ranking

In this section, machine learning is used to support the information flow by aiding in information search. The aim is to ease information access through automated relevance evaluation between a given search topic and text segments. This clinical application is more thoroughly explained in Section 3.1.1, and it is related to the *Attention focusing* and *Summarization* components of the road map for developing HLT (Figure 2.6) and the phase of using narratives (Figure 2.7). The learning task, task-specific performance evaluation measures, and related work are described in Sections 3.1.2, 3.1.3, and 3.1.4, respectively. Sections 3.1.5, 3.1.6, 3.1.7, and 3.1.8 specify the data consisting of Finnish ICU nursing narratives, linguistic processing, learning methods, and performance evaluation setting. Finally, Sections 3.1.9 and 3.1.10 report and summarize the results.

3.1.1 Clinical application

The aim is to support the information flow by developing HLT tools that aid users in finding information relevant for their interests. The clinical application can be visualized as follows: The user defines a topic of interest, and then the tool highlights the relevant parts of the narratives. The user can select the relevance evaluation scale: In the simplest case, the tool has the granularity of relevant versus irrelevant (Figure 3.1). But it can have several degrees (Figure 3.2), up to continuous relevance scaling.

The tool is of clinical significance, because it expedites text browsing. In addition, it does not lose the original context of the highlighted text segments. Further, relevance-degree evaluation can be seen as helpful for writing discharge documents, because segments with the largest topical relevance should summarize the most essential information. Text segments with smaller topical relevance could be used to focus the attention of professionals' in direct care toward potential weak signals of some particular issue in order to examine them more carefully.

The focus is on *topic labeling* and *relevance ranking* of text segments. In other words, text is assumed to be readily divided into segments consisting of one matter or thought, and the task is to evaluatev the topical relevance of each segment.

The topics that are considered include *Breathing*, *Blood circulation*, and *Pain*. These topics are chosen for three reasons. First, they represent crucial intensive care aspects. The emphasis in ICU nursing is on monitoring, assessment and maintenance of breathing, blood circulation, and other vital functions (Ward et al., 2004). Similarly, regular pain assessment and management are crucial but extremely difficult (Sessler et al., 2008; Suominen et al., 2009b). Second, nurses evaluate ICU patients' care needs and nursing intensity regularly, and in this task, they use narratives to support their decision making. In Finland, the evaluations are based on the Oulu Patient Classification (Kaustinen, 1995; Fagerström et al., 2000), which distinguishes six different areas of ICU nursing: planning and co-ordination of care; breathing, circulation and symptoms of diseases; nutrition and medication; personal hygiene and excretion; activity, movement, sleep and rest; as well as teaching, guidance in care, follow-up and emotional support. Within the first area, nurses evaluate breathing, blood circulation and pain. Third, due to the nature of pain evaluation generally, and especially in ICU patients, we assume that pain is documented differently from breathing and blood circulation. This is of interest from a machine learning ability viewpoint.

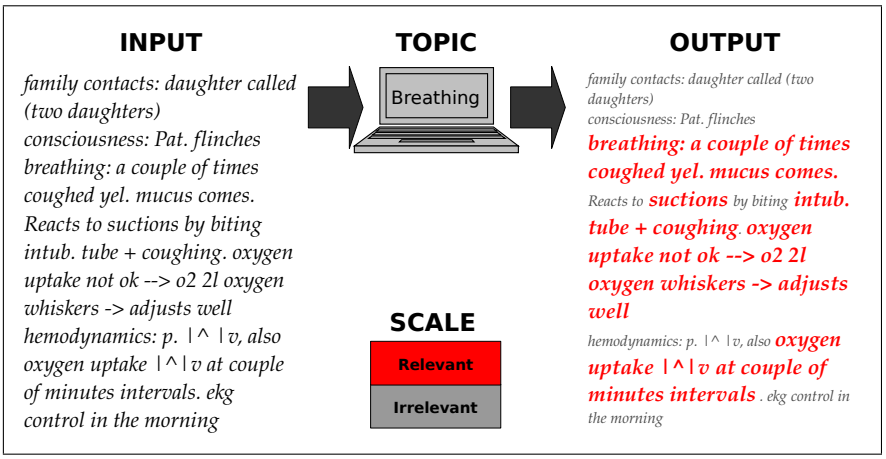


Figure 3.1: *Identification of topically relevant text segments*

3.1.2 Learning task

The focus is on topic labeling and relevance ranking of text segments. In the simplest case (Figure 3.1), the task is to infer for each segment the relevant topics. This belongs to the family of *text classification* tasks in machine learning. In the more advanced case (Figure 3.2), the task is to infer an ordering of the text segments with respect to their relevance to a given topic. This is known as *text ranking* in the machine learning community. The most extreme cases with a continuous relevance-scale belong to the family of *text regression* tasks.

The topic labeling task comprises three *binary classification* tasks, one for each topic. All three topics are considered separately, because one segment can be relevant for many topics. For each segment and topic in turn, the task is to decide, whether or not the segment contains relevant information about the topic. The relevant segments are called *positive instances* and irrelevant *negative instances*.

Our binary classification task can be formalized as follows: The training instances are again $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in \mathcal{X}$ represents the text segment i, $y_i \in \mathcal{Y}$ is the corresponding topic label, $\mathcal{X}$ the *input space* (i.e., if x_i is a feature vector, then $\mathcal{X}$ a space containing all possible feature vectors of the size $|x_i|$), and $\mathcal{Y}$ the *output space*. In binary classification $\mathcal{Y} = \{-1, +1\}$, where -1 (resp. $+1$) is for negative (resp. positive) instances by convention. The task is to learn by observing the training data a function (i.e., a *classifier*) $f : \mathcal{X} \to \mathcal{Y}$ that predicts an output $f(x) \in \{-1, +1\}$ for an input $x \in \mathcal{X}$. Naturally, the goal is to make as few classification mistakes as possible.

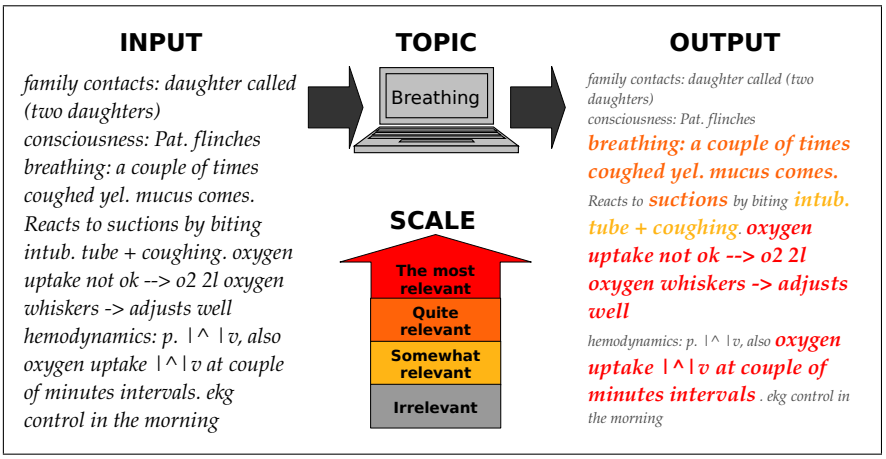


Figure 3.2: *Topic-relevance evaluation*

A text regression task is, in a sense, a generalization of a binary classification task, where the output values $y_i \in \mathbb{R}$ instead of only $\{-1, +1\}$. The output value can be interpreted as a topical relevance-score, with the larger value indicating the stronger topic-relevance. Again, the goal is to have as a good prediction function as possible. Solutions for the classification and ranking tasks can be constructed from regression results. Namely, the binary division between positive and negative instances can be obtained by simply placing a threshold that halves the output. When the evaluation scale contains several relevance-degrees, the ranking of segments is obtained by placing more thresholds.

3.1.3 Performance evaluation measures

We evaluate the performance separately for every topic. Thus, *binary classification measures* and *ranking measures* for one topic at a time are discussed next. The section is based on our study (Suominen et al., 2008b).

Perhaps the most intuitive binary classification performance evaluation measures are *accuracy* and *error*, defined as the proportions of correctly and incorrectly identified instances. Their values belong to an interval [0, 1], but with accuracy, the value one corresponds to the best performance whereas with error, being one minus accuracy, the value zero means the optimum.

However, when the numbers of positive and negative instances are strongly of an unequal size, accuracy and error may give misleading results. For example, let us assume that ninety-five percent of the text segments are irrelevant to the topic of *Breathing*. Then, without any learning from observations, we get a classifier with only a five percent error by labeling all

segments as irrelevant to *Breathing*. Problems also arise, if the proportion of instances left correctly unlabeled (*true negatives*, N_{TN}) is excessively large when compared to the numbers of *true positives* (N_{TP}), *false positives* (N_{FP}) and *false negatives* (N_{FN}). In addition, the number of true negatives may even be unknown, as in the case of applying text classification to identify documents relevant to a query given to a web search engine.

Precision (i.e., the proportion of correctly classified positive instances from all positive instances in the output of the classifier), *recall* (i.e., the proportion of correctly identified positive instances from all instances that should have been identified as positive), and many other classification measures based on these are independent of true negatives. Hence, they are also applicable when the number of true negatives is unknown or excessively large. Both precision and recall get values between zero and one, with higher values indicating the better the performance. As recall can be trivially increased at the cost of decreased precision by assigning more instances to the class, and vice versa, both two measures must be considered together. One possibility is to select a measure, such as F that combines the measures.

In case of F, a weighted harmonic mean of precision and recall is taken. In other words,

$$F = \frac{1}{\beta\frac{1}{\text{precision}} + (1-\beta)\frac{1}{\text{recall}}}, \tag{3.1}$$

where $\beta \in [0, 1]$ is a factor determining the weighting of precision and recall. The most common choice is to weight precision and recall evenly, that is, to select $\beta = 0.5$, which leads to a measure known as $F1$, *F measure* or *balanced F score*. The values of (3.1) are between zero and one, with higher values indicating the better the performance.

All previously mentioned binary classification measures are sensitive to the relative number of positive and negative instances. This class distribution dependence can be problematic if, for example, the distribution is different in the gold standard than it is in the actual application. *AUC* (Hanley and McNeil, 1982) is a measure invariant to class distribution (see, e.g., Fawcett and Flach (2005)). On the other hand, unlike precision, recall and F, it incorporates the number of true negatives. The interpretation of *AUC* is that it is the probability that, given a randomly chosen positive instance and a randomly chosen negative instance, the classifier will correctly distinguish them (Cortes and Mohri, 2004). This can be seen from the following probabilistic definition:

$$AUC = \frac{\sum_{y_i=+1, y_j=-1} \delta(f(x_i) > f(x_j))}{y_+ y_-}, \tag{3.2}$$

where y_+ and y_- are the numbers of positive ($y_i = +1$) and negative ($y_i = -1$) instances in the gold standard and

$$\delta(e) = \begin{cases} 1, & \text{if the expression } e = \text{TRUE} \\ 0, & \text{if the expression } e = \text{FALSE} \end{cases} .$$

The values of *AUC* belong to an interval [0, 1], with larger values indicating better performance.

Note that *AUC* is equivalent to MannWhitney U (see Section 2.4): First the *AUC* value for the label predictions and gold standard is calculated. Then, a vector of the predicted outputs that should have been positive (i.e., elements are those $f(x_i)$ that have $y_i = +1$) and a vector of the predicted outputs that should have been negative are formed, and the MannWhitney U value for those two vectors is computed. Finally, $AUC = U/(y_+y_-)$.

The relation of *AUC* to precision, recall and F is easy to see from a more traditional definition based on computing the *receiver operating characteristic* (ROC) curve first and then calculating the area under the curve. Instead of assuming the system output to be a positive or negative label for each input instance, many classification algorithms produce first a regression output, as explained in Section 3.1.2. The ROC curve refers to plotting the recall (a.k.a. *true positive rate*) at certain levels of the *false positive rate* $N_{FP}/(N_{FP} + N_{TN})$ obtained by varying the values of the threshold parameter between negative and positive instances from the minimum threshold value (i.e., the one resulting all instances to be positives) to the maximum (i.e., the one resulting all instances to be negatives) (Figure 3.3). In addition to the area under the curve (i.e., *AUC*), the ROC curve provides information about the trade-off between recall and false positive rate: The analysis of the curve from left to right corresponds to assigning more instances as positives. That is, improving the performance in terms of recall at the expense of increased false positive rate.

In addition to classification, *AUC* can be applied in regression and ranking tasks as well, because only the pairwise order of the instances with respect to the classification topic (i.e., information defining which of the two instances is larger), is needed. However, there are ranking task-specific measures too.

Established ranking measures are based on evaluating the extent to which the output ranking agrees with the gold standard ranking, typically by determining their correlation. Consequently Kendall's τ and other general measures of rank correlation applicable (see, e.g., Siegel and Castellan (1988, pp. 235–254) and Kendall and Gibbons (1990) for further information about the rank correlation coefficients). However, in evaluating ranking performance, it is essential to take into account the possibility of tied ranks, where two or more instances have the same rank in either output ranking or the gold standard ranking. In order to address the issue of tied ranks,

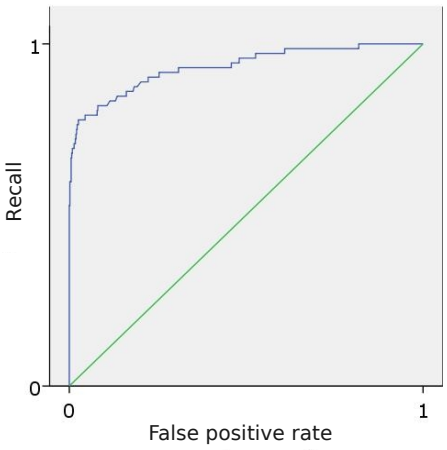


Figure 3.3: *ROC curve corresponding to the content expert E_2 and classifier $C(E_2)$ in the topic of Blood circulation*

tie-corrected versions of rank correlation measures have been defined. In particular Kendall's τ_b (Kendall and Gibbons, 1990, p. 40), a tie-corrected rank correlation measure, has gained popularity in evaluating machine learning performance. It is formally defined as

$$\tau_b = \frac{g(f(X), Y)}{\sqrt{g(f(X), f(X))g(Y, Y)}}, \tag{3.3}$$

where

$$g(f(X), Y) = \sum_{i,j=1}^{m} \text{sign}(f(x_i) - f(x_j))\text{sign}(y_i - y_j),$$

$f(X)$ is a vector containing the output ranking of the input X of m feature vectors x_i, $Y = (y_1, \ldots, y_m)$ is the respective gold standard, and

$$\text{sign}(e) = \begin{cases} -1, \text{ if the expression } e < 0 \\ 0, \text{ if the expression } e = 0 \\ 1, \text{ if the expression } e > 0 \end{cases}.$$

When the compared rankings are identical, $\tau_b = 1$; when one ranking is reverse of the other, $\tau_b = -1$; otherwise, $\tau_b \in (1, 1)$.

3.1.4 Related work

Text classification and ranking are established areas in machine learning research. Trends[1] in their recent health care applications are an extraction of certain specific information from patient records and identification of patients with a given diagnosis (Table 3.1). In addition, applications to generate alarms and ease information search or browsing are described. Of these applications, diagnosis coding can be seen as a direct text classification task whereas others apply classification and ranking among other HLT techniques. For example, information extraction applications also need methods to automatically distinguish the relevant text segment.

Regardless of the wealth of studies with promising results, clinical applications are still relatively rare (Collier et al., 2006), and little research exists on the application domain of nursing narratives (Bakken et al., 2005). For example, from the fifteen studies summarized in Table 3.1, only four report an application that is in clinical use or trial phase, and none of the studies is specifically tailored for nursing narratives. None of the studies address Finnish. The investigation of ICU narratives is explicitly mentioned in two studies.

[1]Hanna Suominen and Heljä Lundgrén-Laine conducted a PubMed search on 2007 November 12. We utilized the MeSH hierarchy with our query "patient records AND (natural language processing OR language technology OR decision support systems, clinical OR decision support OR decision making) AND (narrative* OR note* OR text*)" and limits published in the last 2 years, Humans, English. Our search returned 42 documents, of which we both considered the fifteen cited in Table 3.1 as related to health care applications of human language technology.

Table 3.1: *Literature review results*

Publication	**Aim**	**Narratives**	**ICU narratives**	**State**
Currie et al. (2006)	Extract factors related to cancer	8,664 US patient records	No	Use
Friedlin and McDonald (2006)	Extract family history data	1,337 US admission documents	No	Research
Meystre and Haug (2006a)	Extract medical problems	160 US patient records	No	Research
Meystre and Haug (2006b)	Extract medical problems	105 US patient records	Yes	Trial
Shah and Martinez (2006)	Extract daily medication doses	One million UK prescriptions	No	Research
Turchin et al. (2006)	Extract hypertension treatment data	600 US physician narratives	No	Research
Zeng et al. (2006)	Extract factors related to asthma	150 US discharge documents	No	Research
Zhou et al. (2006)	Extract medical terms	888,357 US patient records	No	Research
Pentz et al. (2007)	Extract adverse events	316 US patient records	Yes	Research
Pakhomov et al. (2006)	Identify patients with certain diagnoses	23 million US patient records	Unknown	Use
Hripcsak et al. (2007)	Identify patients with pneumonia	11,698 US discharge documents	No	Research
Pakhomov et al. (2007b)	Identify patients with angina pectoris	892 US physician narratives	Unknown	Research
Pakhomov et al. (2007a)	Identify patients with heart failure	2,904 US patient records	No	Research
Erdal and Kamal (2006)	Ease browsing	Unknown	No	Research
Gundlapalli et al. (2007)	Generate alarms for surveillance	US radiology tests and results	No	Use

3.1.5 Patient record data

The Finnish patient record data considered in this section is collected in 2001 from fifteen ICUs for adults. From every unit, the records of three patients have been chosen; the inclusion criterion is an alphabetized order on the collection day. The records include the patient's admission situation and daily nursing notes for one 24 hour period. The total number of records is 43, as two units have sent only two patient records instead of the three requested. The data is divided into training and test sets in which the narrative statements from the individual documents are aggregated, but the recognition of individual patients is impossible.

The text style in the nursing notes varies in length.

- *Virtsa tummaa.* [*Urine dark.*],
- *Äiti soittanut.* [*Mother called.*],
- *p ok.* [*P OK.*], and
- *Sao2 < 90.* [*Sao2 < 90.*]

illustrate short sentences containing only one matter, whereas

- *Valuskelee herkästi verta (kääntämällä käsivarren sidoksiin verta) katsottu ihonsiirros, hematoomainen, saatu hematooma osin pois -> päälle rasvataitos + kompr. kostea vety(?) taitos.*
  [*Is bleeding easily (blood to arm bandages when turning) the skin graft checked, haematomic, the haematoma partially removed -> a paraffin gauze dressing on top + a compr. moist hydrogen(?) dressing.*] and
- *Pulssi epätasaista anest. lääkäri käynyt katsomassa ekg nauhaa (otettu ennen leikkausta) sinusextroja, ekg kontrolli huomenna aamulla, käännöllä potilas muuttui siniseksi ja saturaatio putosi 50, 100% hapella ambutuksella tilanne korjautui sedaatiota nostettu.*
  [*Pulse uneven anaesthetist checked the EKG tape (measured before the op) extra systoles, EKG control tomorrow morning, became bluish when turning him around and the saturation decreased to 50, with 100% oxygenation with ambutus the situation improved, sedation increased.*]

are example sentences with multiple matters.

In order to normalize the style of the nursing notes, a content expert is used to divide the text into segments consisting of one matter or thought. After manual segmentation, the data contains $1,363$ segments, with the average length of 3.7 words.

Three content experts (E_1, E_2, and E_3) are asked to annotate the patient data. Their task is to evaluate the content of the each text segment and

mark those related to the topics of *Breathing*, *Blood circulation*, and *Pain*. All three annotators are registered nurses and specialists in nursing documentation. E_1 and E_2 have a long clinical experience from ICUs, whereas E_3 is an academic senior nursing science researcher. The annotation guideline is to label a segment if it includes relevant information about the given topic.

The annotations are used directly, as they are in the topic labeling task (Figure 3.1). In order to evaluate the performance of the topic classifiers in supporting health communication, we obtain both classifiers' performance with respect to the content experts as well as content expert against other content expert comparisons.

For the text ranking task (Figure 3.2), the gold standard ranking for each text segment is obtained by calculating the number of content experts that have labeled it as belonging to the given topic. The resulting gold standard associates each text segment with a rank from zero to three and the magnitude of this rank is assumed to reflect topic-relevance. The assumption is based on seeing the disagreements between the experts as a valuable resource that reflects different experiences, knowledge and expertise areas.

3.1.6 Linguistic processing

Topic labeling and relevance ranking are our first machine learning applications for clinical text. Hence, our aim is to assess feasibility of the machine learning approach. This justifies our focus on comparing content experts' opinions and machine learning outputs with very little investment in linguistic processing.

Before the machine learning phase, we perform the following linguistic processing: First, to unify the data, punctuation marks and special characters are separated from words with spaces and all letters are converted to lower case. Then, in order to reduce different inflection forms of the words, the data is linguistically processed using the Snowball stemmer for Finnish (Porter and Boulton, 2006).

3.1.7 Learning methods

The automated learning tasks are performed by using the *regularized least-squares* (RLS) *algorithm* (see, e.g., Rifkin (2002) and Poggio and Smale (2003)), a kernel-based learning algorithm that is also known as the least-squares support vector machine or the subset of regressors method (Suykens and Vandewalle, 1999). RLS has been shown to have a state-of-the-art performance in classification and regression, and it has been successfully modified for other problems such as ranking (see, e.g., Pahikkala et al. (2009)).

To define the algorithm formally, let the set of training instances be $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in \mathcal{X}$ describe the text segments, $\mathcal{X}$ is a space of bag-of-words vectors, and y_i is the gold standard output. For topic labeling, $y_i \in \{-1, +1\}$ and for relevance ranking $y_i \in \{0, 1, 2, 3\}$.

The RLS algorithm can be considered as a special case of the following regularization problem known as the Tikhonov regularization:

$$\min_f \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_k^2, \tag{3.4}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a function defining the output of the algorithm, $\lambda \in \mathbb{R}_+$ 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 kernel function $k(x_i, x_j)$ evaluates the similarity of instances x_i, $x_j \in \mathcal{X}$ and the reproducing kernel Hilbert space basically guarantees that k exists and distances between input vectors can be measured (i.e., $k(x_i, x_j)$ exists for all x_i, $x_j \in X$ and $\|\cdot\|_k$ is well-defined). The least-squares loss function $(y_i - f(x_i))^2$ is used to select a classifier that performs well with training instances and the regularizer $\lambda\|f\|_k^2$ controls overfitting. By the Representer Theorem (Schölkopf et al., 2001), minimizers of equation (3.4) have the form

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i), \tag{3.5}$$

where $\alpha_i \in \mathbb{R}$.

The RLS algorithm is trained for each of the topics separately. The kernel function is the cosine of the word feature vectors, that is,

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

3.1.8 Performance evaluation

Performance evaluation contains three aspects: the content experts' agreement on topic labeling, the performance of the topic labeling application, and the performance of the text ranking application. All the statistical analysis is performed with SPSS 11.0 for Windows.

Content experts' agreement is assessed separately for each topic by using *Cohen's* κ (Cohen, 1960). Cohen's κ is an inter-annotator agreement measure that considers the two annotators being compared as equally competent to make judgments, places no restriction on the distribution of judgments over topics, and takes into account that a certain amount of agreement is to be expected by chance. Formally, it is defined as

$$\kappa = \frac{A - A_c}{1 - A_c},$$

Table 3.2: *Inter-annotator agreement (Cohen's κ and 95 percent confidence interval) of the content experts E_1, E_2, and E_3 on the topics of Breathing, Blood circulation, and Pain with the data of 1,363 text segments*

Reprint from Paper II

Comparison	Breathing	Blood circulation	Pain
E_1 vs. E_2	0.73 (0.68–0.78)	0.89 (0.85–0.92)	0.88 (0.82–0.94)
E_1 vs. E_3	0.67 (0.62–0.72)	0.81 (0.77–0.86)	0.79 (0.73–0.86)
E_2 vs. E_3	0.85 (0.82–0.89)	0.87 (0.83–0.90)	0.76 (0.69–0.83)

where A is the proportion of times the annotators agree and A_c the proportion of times they would be expected to agree by chance. The values of $\kappa < 0$ for poor agreement, and values between zero and one denote agreement; the closer to one, the better the agreement.

For evaluating the automated classifier in the topic labeling task, its output is compared with the content experts' opinions by using *AUC* (3.2). Text ranking performance evaluation is performed with Kendall's τ_b (3.3). The same measures are also in use for model selection (λ and threshold values). For performance evaluation and model selection, the data is halved: The first 708 text segments are used for training the classifier and the remaining 655 segments are used for testing. The division is done so that segments from one patient record belong only in one of the two files. Leave-one-out cross-validation on the training set is chosen for model selection. In total, nine automated classifiers are trained, that is, a distinct classifier for each annotator–topic pair. The number of text ranking applications is three, one for each topic.

3.1.9 Evaluation results

This section describes the empirical results, and it is based on Papers II and III. However, I have performed some additional analysis for the dissertation.

Gold standard topic labeling

The content experts labeled approximately twenty, fifteen and six percent of the 1,363 text segments as belonging to the topics of *Breathing*, *Blood circulation* and *Pain*, respectively. The pair-wise comparisons between the annotations showed that text segments related to *Blood circulation* were selected quite similarly, whereas the number of disagreements was larger in *Breathing* and *Pain* (Table 3.2).

The performance of the algorithm was tested in two ways. First, the learn-

Table 3.3: *Classification performance (AUC and* 95 *percent confidence interval) in the topics of Breathing, Blood circulation, and Pain with separate training (*708 *text segments) and test sets (*655 *text segments)*

Extended from Paper II

$C(E_i)$: the classifier trained with the data labeled by the content expert E_i

Breathing

	E_1	E_2	E_3
$C(E_1)$	0.86 (0.82–0.90)	0.74 (0.69–0.79)	0.72 (0.68–0.77)
$C(E_2)$	0.83 (0.79–0.88)	0.88 (0.85–0.91)	0.86 (0.83–0.89)
$C(E_3)$	0.84 (0.80–0.88)	0.88 (0.84–0.91)	0.87 (0.84–0.91)

Blood circulation

	E_1	E_2	E_3
$C(E_1)$	0.89 (0.84–0.93)	0.93 (0.89–0.97)	0.91 (0.87–0.95)
$C(E_2)$	0.88 (0.83–0.93)	0.93 (0.90–0.97)	0.91 (0.86–0.95)
$C(E_3)$	0.89 (0.84–0.93)	0.93 (0.90–0.97)	0.91 (0.86–0.95)

Pain

	E_1	E_2	E_3
$C(E_1)$	0.71 (0.61–0.80)	0.81 (0.72–0.90)	0.72 (0.63–0.81)
$C(E_2)$	0.71 (0.61–0.80)	0.81 (0.73–0.89)	0.71 (0.62–0.80)
$C(E_3)$	0.67 (0.56–0.78)	0.77 (0.66–0.87)	0.71 (0.61–0.80)

ing ability was evaluated by comparing the output of the algorithm with the opinions of the same content expert who had also labeled the training data. Second, generalization capabilities of the algorithm were evaluated by comparisons not only with the same content expert but also with the other two experts. In all evaluations, separate training and test sets were used, as explained in Section 3.1.8.

The results of the first evaluation showed that the algorithm was able to learn the classification task (Table 3.3). The most encouraging results were achieved in the topics of *Breathing* and *Blood circulation*. The topic of *Pain* was more challenging to learn, as expected; in this topic, the algorithm had the smallest amount of positive instances for learning and it seemed to be difficult for content experts too. In the topics of *Breathing* and *Blood circulation*, the performance of the algorithm was on a similar level regardless of whose opinions it was trained with.

Table 3.4: *Number of text segments in the gold standard ranking*

Rank	Breathing		Blood circulation		Pain	
	Training	Testing	Training	Testing	Training	Testing
0	572	462	564	566	646	609
1	28	31	26	15	14	14
2	35	60	23	9	7	10
3	73	102	95	65	41	22
1–3	136	193	144	89	62	46

These results illustrate the extensive content of nursing narratives and consequent difficulties in searching information. In addition, they justify the selected performance evaluation measure.

Text labeling performance

The results of the second evaluation gave evidence of the relatively good generalization capability of the algorithm across the content experts' opinions; the differences in the *AUC* values were relatively small regardless whose labeling was used in training or testing (Table 3.3). This was particularly emphasized in the topic of *Blood circulation*. In terms of the average difference in the *AUC* value compared to the situation, where both the training and testing data were labeled by the same content expert, the decrease was negligible in the topics of *Blood circulation* and *Pain* (i.e., 0.00 and 0.01, respectively). The decrease was somewhat larger in the topic of *Breathing* (i.e., 0.06), where the content experts' disagreement was also the largest (Table 3.2).

Gold standard relevance ranking

The gold standard ranking contained considerably more text segments with relevance rank zero than those with larger ranks (Table 3.4). In particular, the data had very few instances with relevance ranks of one and two.

In the gold standard ranking, the topics of *Breathing*, *Blood circulation*, and *Pain* had similar general characteristics of the ranks. For example, in the topic of *Breathing*, text segments with the rank three were related to issues such as oxygenation and the use of different kinds of respiration devices, whereas segments with lower ranks described mucus in patients' lungs, coughing, and issues related to pleural tubes. In the topic of *Blood circulation*, segments with the highest rank were mostly about the progression of various measurement values, such as pulse or blood pressure; those with two or one described issues such as body temperature and bluish skin

Table 3.5: *Ranking performance (Kendall's τ_b and* 95 *percent confidence interval) in the topics of Breathing, Blood circulation, and Pain with separate training (*708 *text segments) and test sets (*655 *text segments)*

Reprint form Paper III

Breathing	**Blood circulation**	**Pain**
0.62 (0.56–0.68)	0.69 (0.61–0.76)	0.44 (0.30–0.59)

shade. In the topic of *Pain*, segments with the rank three commonly contained keywords such as *kipu* [*pain*] or *särky* [*ache*], whereas segments with rank one or zero included neither the word *kipu* [*pain*] nor its derivatives or synonyms. Segments with the pain rank one or two covered implicit pain indicators such as the quality of sleep and reactions to nursing interventions.

In conclusion, the gold standard ranking seems justified; segments with a rank three had the most evident notes about the topic, and the connection to the topic became more implicit, when the relevance rank decreased. However, the algorithm has extremely few observations with ranks one and two for learning.

Relevance ranking performance

The best ranking performance was in the topics of *Blood circulation* and *Breathing*(Table 3.5). The learning task related to the topic of *Pain* was again more challenging. The main reason for this is the considerably smaller number of relevant training instances; altogether only 62 out of 708 training instances got the pain rank larger than zero in the gold standard ranking. In the other two topics, where the text ranking application performed better, the algorithm had more than twice the amount of relevant instances for learning. As an example, all text segments including the word *päänsärky* [*headache*] were ranked mistakenly to the pain rank zero instead of the correct rank three because the training set did not contain this word. Evidence of the importance of linguistic processing was provided in preliminary experiments; the performance of the RLS algorithm was better with the stemmed data than with the original data.

Finally, to illustrate the potential of text ranking for summarizing the most essential information, let us consider a need to quickly build an overview about issues related to blood circulation. In this case, the user could select the sensitivity level to be, for example, fifty or one-hundred segments, and then the ranking application would return the corresponding number of text segments in ascending order of relevance (Figure 3.4). In the gold standard ranking of 655 segments in the test set, the number of segments with ranks zero, one, two, and three were 566, 15, 9, and 65, respectively (Table 3.4).

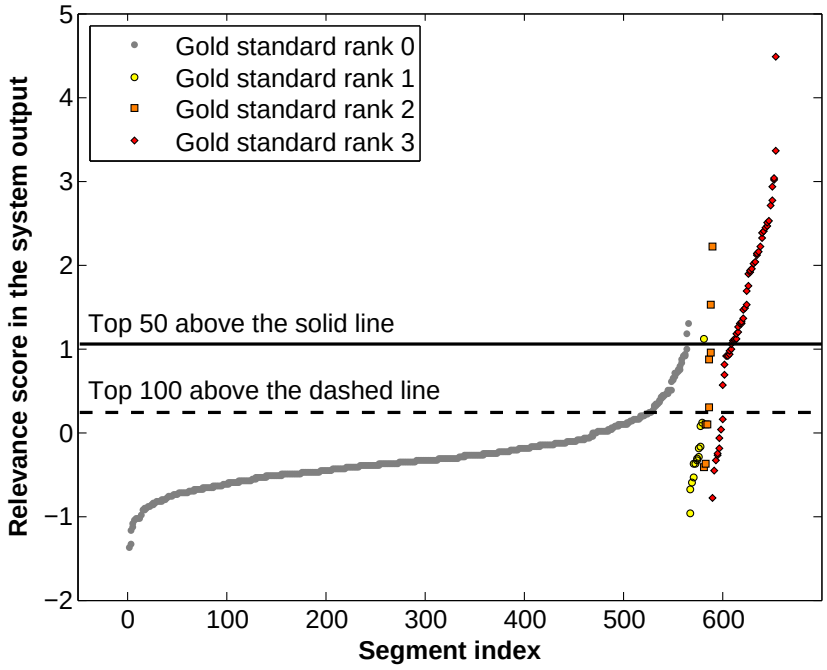


Figure 3.4: *Tested sensitivity level in the topic of Blood circulation*

With the sensitivity level fifty, a great majority of returned text segments (i.e., 45 segments) belonged to the blood circulation rank three in the gold standard ranking. The number of returned segments with ranks one and two in the gold standard were one and two, respectively. Two returned segments (*tilanne stabiili* [*situation stable*] and *nostettu infuusiota* [*lift in infusion*]) were considered irrelevant in the gold standard.

With the sensitivity level of one-hundred text segments, 55 segments with the rank three in the gold standard were returned. The number of returned segments with ranks one and two in the gold standard were one and five, respectively. However, the number of segments that were considered irrelevant in the gold standard increased to 39. If all segments with ranks one, two, or three in the gold standard had been returned, the number of topically irrelevant segments should have been only eleven in the system output of one-hundred segments.

In conclusion, the results of text ranking are promising. Learning was already possible with a relatively small number of topically relevant segments, and the text summarization example showed that the application was able to distinguish the most relevant text segments.

3.1.10 Summary

The section has discussed binary classification and regression, and their application to automated topic labeling and relevance ranking of patient records. The results indicate a success of machine learning in the applications, but with better performance in the topics of *Blood circulation* and *Breathing* than in the topic of *Pain*.

The performance differences are used to guide future research and also my dissertation. First, the most crucial consideration for future research is to divide text into segments automatically.

Second, the amount of topically relevant training instances had a substantial effect on learning ability. Consequently, more data is used in Section 3.2.

Third, performance differences may indicate topical variability both in terms of the textual content and learning task difficulty. Thus, more topics are considered in Section 3.2. However, these topics (*Breathing*, *Hemodynamics*, *Consciousness*, *Relatives*, *Diuresis*, and *Other*) do not explicitly include the interesting, but challenging, topic of *Pain*. We have begun a more detailed study of the topic in Suominen et al. (2009b). This study reports how a group of ten nursing professionals were supervised to annotate a set of 1,548 daily nursing notes, and based on these annotations, builds a gold standard. The annotation aspects include the amount and writing style of pain-related notes, pain intensity, and given pain care. The conclusion is that more than half of the analyzed] documents contains information relevant for patients' pain status or medication but it is usually expressed indirectly. Although annotators' pain intensity evaluations diverged, the substantial amount of pain-related notes encourages developing computational tools for pain assessment.

Finally, the section gave empirical evidence of the importance of linguistic processing to reduce the data sparseness. More sophisticated linguistic processing is likely to contribute to the learning performance generally, and in particular in case of limited amount of data; it could enable recognizing *kipu* [*pain*], *särky* [*ache*] and other topical keywords from compounds (e.g., *päänsärky* [*headache*], *kipulääke* or *särkylääke* [*painkiller*], *päänsärkylääke* [*headache painkiller*], and *kipukynnys* [*pain threshold*], derivatives (e.g., *kivulias* [*painful*] and *kivuton* [*painless*]), as well as numerous inflection forms in Finnish (e.g., *kivun*, *kipuna*, *kipua*, *kivuksi*, and *kivusta*).

3.2 Topic segmentation and labeling

This section continues the study with information search applications discussed in the previous section. Now, not only topically relevant segments are labeled, but the application performs the automated text division into

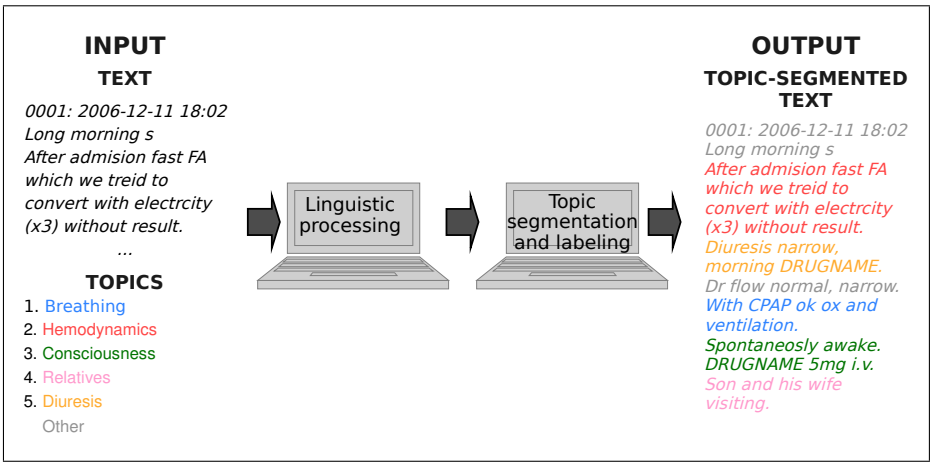


Figure 3.5: *Topical segmentation and labeling*

topical segments too. Furthermore, it addresses domain-tailored linguistic processing, more topics and larger, but more focused, data. The application is again related to the *Attention focusing* and *Summarization* components of the road map (Figure 2.6) and the phase of using narratives (Figure 2.7). The clinical application, learning task, task-specific performance evaluation measures, and related work are described in Sections 3.2.1, 3.2.2, 3.2.3, and 3.2.4, respectively. Sections 3.2.5, 3.2.6, 3.2.7, and 3.2.8 specify the data consisting of Finnish ICU nursing narratives, linguistic processing, learning methods, and performance evaluation setting. Finally, Sections 3.2.9 and 3.2.10 report and summarize the results. The section is based on Paper IV.

3.2.1 Clinical application

The clinical application for information search is visualized in Figure 3.5. Again, the user can specify topics of interest, and then the tool highlights the narratives according to these topics. In order to serve clinical needs for browsing nursing narratives, the topics that are used in the method development are the most common ones of these records (see Chapter 2 and Section 3.2.5): *Breathing* (including both *Breathing* and *Oxygenation*), *Hemodynamics*, *Consciousness*, *Relatives*, and *Diuresis*. To enable fast browsing, the application assigns for each word the most relevant topic only. However, this approach loses the aspect of one text segment being relevant for many topics. Text that is irrelevant to all topics is left unlabeled.

3.2.2 Learning task

In *text segmentation*, the aim is to divide text into smaller semantically coherent units by placing boundaries between its elements. The elements can be individual words, but also sentences, paragraphs or other larger text entities. In this dissertation, the element is a word and semantic coherence means topical similarity. That is, the goal is automatic division of the input text into topically coherent units by placing topic change boundaries between words. This is known as *topic segmentation*. Not only topic change boundaries are to be assigned but also the topics of the resulting segments. Hence, the application is called *topic segmentation and labeling*.

Topic segmentation and labeling belongs to the family of *multi-class classification* in the machine learning task taxonomy. Topic segmentation alone can be viewed as a binary classification task, where the choice is whether or not to place a topic change boundary between two text elements of the input text. If topic labeling is also considered, the task can be seen as multi-class classification, where for each input word, one topic label is to be assigned from a set of possible topics.

3.2.3 Performance evaluation measures

Evaluation of topic segmentation and labeling quality combines the criteria for the topic segmentation task and topic labeling task. The section is based on Suominen et al. (2008b) and Paper VI.

Because topic segmentation and labeling can be treated as text classification, many classification measures can also be chosen to be used here. The use of accuracy, error, precision and recall measures is especially common (Hirschman and Mani, 2003, p. 416). If topic segmentation is considered alone, then each segment boundary produced by the segmentation application is interpreted as correctly (true positive) or incorrectly (false positive) placed. Undetected boundaries and potential segment breaks that are correctly left without a boundary represent false negatives and true negatives, respectively. Accuracy, error, precision, recall and other measures based on the proportions of N_{TP}, N_{FP}, N_{TN}, and N_{FN} can thus be defined as previously. However, these measures are not sensitive to situations where the segment boundary is almost correctly placed, but not exactly. As a result, topic segmentation specific measures have been developed.

The P_k *measure* (Beeferman et al., 1999) is a popular topic segmentation specific performance evaluation measure (see Pevzner and Hearst (2002)). P_k is a special case of the P_D *measure* (Beeferman et al., 1999), which is the probability that two elements drawn randomly from the input are correctly identified as either belonging or not to the same segment in the system output with respect to the gold standard. P_D contains a distance probability

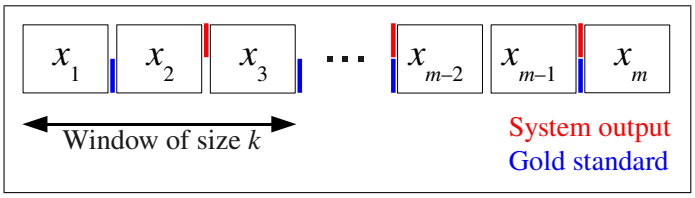


Figure 3.6: *Window sweeping across the input*

distribution over the set of possible distances between two randomly chosen text elements, and in P_k, this function is fixed so that it produces P_k to the probability that a randomly chosen pair of text elements that are k text units apart is assigned to the same segment in either the system output or in the gold standard but not in both. The computation of P_k can be viewed as sweeping a window with a fixed width k across the input text $[x_1, x_2, , x_2, \ldots, x_m]$ (Figure 3.6). Following the explanation given by Pevzner and Hearst (2002), P_k is calculated by determining for each window location whether the outermost elements in the window are incorrectly assigned to the same segment or to different segments in the system output. When observing an inconsistency, the value of P_k is increased by one. Finally, the value of P_k is normalized to be a probabilistic measure, that is, to belong to an interval [0, 1], by dividing by the number of comparisons taken $m - k$, and for P_k, smaller values indicate better performance. In practice, k is set to be half of the average segment size in the gold standard.

Several properties of the P_k measure, however, have been criticized (Pevzner and Hearst, 2002). First, a topic segmentation performance evaluation measure should also take into account almost correctly placed boundaries; P_k achieves the goal in most cases only with false positives, which often causes them be penalized less heavily than false negatives. It can even penalize slightly erroneous boundary placements more than pure false positives of equal magnitude. Second, P_k may leave some topic segmentation errors without penalization, and third, P_k often misses or under-penalizes mistakes in small segments.

To address these drawbacks, another commonly used topic segmentation performance evaluation measure, *WindowDiff* (Pevzner and Hearst, 2002) has been developed. As P_k, WindowDiff can also be viewed as sweeping a window with a fixed width k across the input text $[x_1, x_2, , x_2, \ldots, x_m]$, but for each window position this measure compares the number of boundaries within the window in the system output with the respective number in the gold standard. The measure penalizes the segmentation application by one whenever these numbers differ. Finally, the value of WindowDiff is normalized by dividing by the number of comparisons taken $m - k$. As with

P_k, the value of WindowDiff belongs to an interval [0, 1], and smaller values indicate better performance.

Also the WindowDiff measure has been criticized (Georgescul et al., 2006): It is upbraided for weighting with the same normalization constant irrespectively of having a larger or smaller number of segment boundaries than supposed in the system output. This causes false negatives to be less penalized than false positives even though the weighting should be equal for both of these error types.

Another alternative to evaluate the quality of topic segmentation and labeling application is to consider it as multi-class classification and use classification evaluation measures with *macro-* or *micro-averaging*. This is to get a joint performance estimate for all topics. By comparing the system output with the gold standard, N_{TP}, N_{FP}, N_{TN}, and N_{FN} are defined separately for each topic. Then, the overall performance is defined through macro- or micro-averaging.

Let us consider F (3.1) as an example. The macro-averaged measure is achieved simply by calculating the average of the topic-wise F values. In contrast, micro-averaged F evaluates the performance by computing the measure value based on the global perspective of all binary decisions made. If n_t is the number of topics, then the adjusted numbers of true and false positives and negatives are $TP' = \sum_{i=1}^{n_t} TP_i$, $FP' = \sum_{i=1}^{n_t} FP_i$ and $FN' = \sum_{i=1}^{n_t} FN_i$. Then the micro-averaged precision and recall are calculated by using these adjusted numbers, and finally, micro-averaged F is computed from these. Macro-averaged F emphasizes the significance of performing well on all topics, including relatively rare ones, whilst micro-averaged F weights each code assignment decision equally.

However, this approach sustains the previously discussed problem with almost correctly placed segment boundaries. To address both topic segmentation and labeling aspects, an evaluation setting can cover simultaneous studying of two measures. For example, micro-averaged accuracy could be chosen for the the multi-class classification task and WindowDiff for the binary segment boundary assignment task. Another alternative is to use information extraction measures that incorporate partially correct outputs (see, e.g., Chinchor and Sundheim (1993); Hirschman et al. (2005)). However, they are less conventional in topic segmentation and labeling tasks, and may require manual human judging.

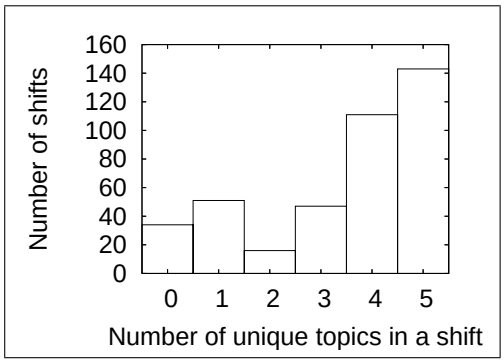
3.2.4 Related work

Topic segmentation and labeling is a well-studied HLT problem. In clinical applications, it has been applied to English medical narratives from radiology and urology departments (Cho et al., 2003). The method relies strongly on hard-coded rules for headings, linguistic cues and lexical patterns seen within training instances. Another example is the temporal order analysis of English medical discharge summaries (Bramsen et al., 2006). This application uses first a statistical parser to segment the sentences into clauses and then two supervised classifiers to predict the segment boundaries and assign for every segment pair their time-wise order. Finally, conditional random fields (Lafferty et al., 2001) have been applied to topic segmentation and labeling of typed English medical dictations (Jancsary and Matiasek, 2008). The aim is to ease the job of the typist through automated text structuring. The method is supervised.

Our topic segmentation and labeling task restricts the applicability of existing machine learning methods. First, the method must produce a segmentation given by the topics of interest that are declared in advance. With unsupervised methods we can gain a free, *ad hoc* choice of topics and freedom from manually annotating a large amount of training data. Unsupervised methods commonly analyse the similarity (e.g., first uses of words, word co-occurrence, repetition or semantic relations) of text before and after a proposed segment boundary (see, e.g., Hearst (1997); Ferret (2002)). A sudden drop in the similarity value indicates a likely change in topic. However, these methods do not typically allow specifying the topics in advance, and topic-sensitive methods tend to be supervised. The supervised methods are often based on probabilistic models for sequence labeling, for instance, on conditional random fields or *hidden Markov models* (HMMs) (Rabiner, 1989; Yamron et al., 1998; Blei and Moreno, 2001; Gruber et al., 2007); the approach is natural because segmentation is inherently given by the assigned topic labels.

Second, we need a method for text, where segments are very short and almost every document contains relevant information about all topics of interest. Existing unsupervised topic segmentation methods require considerably longer segments (e.g., the TextTiling method (Hearst, 1997) searches for topic boundaries between contexts of two hundred tokens whereas in our data, the average segment size is eighteen tokens) and those specifically designed for short segments (see, e.g., Ponte and Croft (1997); Chang and Lee (2003) for methods using the likely topic length and techniques similar to query expansion in information extraction) do not consider pre-specified topics.

Consequently, we aim at a minimally supervised, probabilistic model for sequence labeling and compare performance of supervised and unsupervised



Reprint from Suominen et al. (2008a)

Figure 3.7: *Number of topics (Breathing, Hemodynamics, Consciousness, Relatives, and Diuresis) discussed per shift*

methods with varying amounts of training data. The method referred to hereafter as *LSA-HMM* is based on HMMs because incorporating the likely topic length into conditional random fields is difficult (Sarawagi and Cohen, 2005).

3.2.5 Patient record data

Daily nursing notes described in Section 2.3 are used. To summarize their characteristics, the data set includes 1.2 million words altogether (including punctuation), which are stored into patient- and nursing shift-specific documents. On average, each shift contains about 70 words. The writing style is telegraphic and highly specific with a substantial amount of unit-specific practices and terminology. Approximately half of the shifts are structured by using colon-separated, but non-standardized headings, as *Hemodynamics*, *HAEMODYNAMICS*, *H e m o d*, and *Homedynamics*.

To create topic-annotated data for experiments, three shifts per patient are selected from the records of 135 patients. The admission order is applied in the selection, so that the first patients are chosen. We segment and label the records manually by using the Knowtator tool (Ogren, 2006) of Protégé 3.3.1 Ontology Editor and Knowledge Acquisition System[2]. Irrelevant parts are given the label *Other*. In the manually annotated data, the average shift length is 78 words while the average segment length is only 18 words. Typically, all or almost all five topics are discussed within one shift although 34 shifts contain none of the topics (Figure 3.7).

[2]See `http://protege.stanford.edu/` [cited 2009 August 1].

3.2.6 Linguistic processing

To reduce data sparseness caused by the highly inflective nature of Finnish, we lemmatize the data using a tailored version of the FinTWOL Finnish morphological analyser[3] (Koskenniemi, 1983). This work has been performed in close collaboration with Lingsoft.

We have extended the vocabulary of the analyser with approximately 3,500 clinical terms. For every word analysed by FinTWOL, we used the first lemma given. For example, a Finnish word *haavan* [*wound's*] has the candidate lemmas *haapa* [*aspen*] and *haava* [*wound*]. This reduces the data sparseness and for our statistical methods, it does not matter that we use the wrong meaning (*aspen*) as long as the same mapping is selected systematically. For words not in the FinTWOL lexicon, we preserve the original spelling. This linguistic processing also separates punctuation marks and special characters from words with spaces and converts words, other than those intentionally capitalized, to lower case.

3.2.7 Learning methods

Three methods are compared: an unsupervised keyword search, HMM, and LSA-HMM. These methods are described with more details in Paper IV.

Keyword search

The keyword search inherently resembles the structure of our data. It searches for the occurrence of the five topic keywords (*breathing* etc.) and assigns each word to a topic corresponding to the previous seen keyword. The topic label is given the initial value *other* at the start of each shift.

HMM

Let us denote the topics of interest as $q_i, i \in \{1, \ldots, N_q\}$. Our topic segmentation and labeling task is to infer for the input word sequence $w = [w(1) \ldots w(T)]$ the topic sequence $q = [q(1) \ldots q(T)]$, where $w(t)$ belongs to the vocabulary $\{w_1, \ldots, w_{N_w}\}$ of N_w unique words and $q(t) \in \{q_1, \ldots, q_{N_q}\}$ for all $t \in \{1, \ldots, T\}$. This can be modeled with a first-order HMM, where w is observed and q is hidden, a particular hidden variable $q(t)$ only depends on the previous hidden state $q(t-1)$, an observed variable $w(t)$ is only dependent on the value of the hidden variable $q(t)$, and the random variable describing the start of the chain is uniformly distributed (Figure 3.8). Formally, if $\mathcal{Q}$ is the space of all hidden state sequences, we infer the best q by

[3] See `http://www.lingsoft.fi/` [cited 2009 August 1].

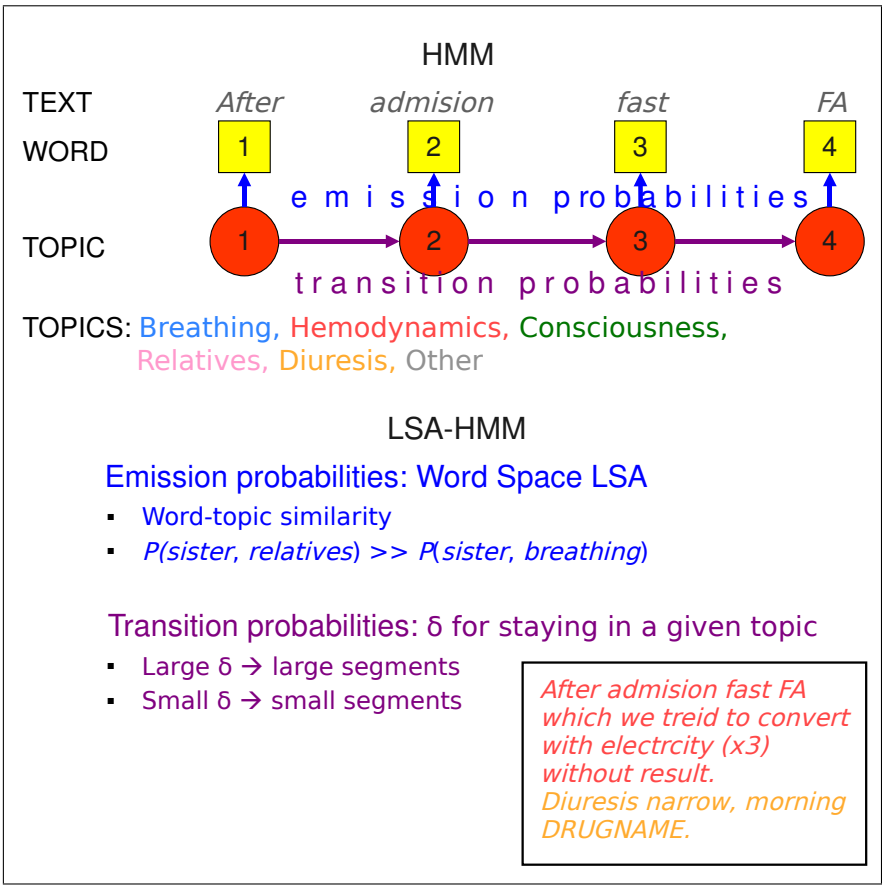


Figure 3.8: *HMM and LSA-HMM methods*

solving

$$\arg\max_{q\in\mathcal{Q}} P(w(1)|q(1)) \prod_{t=2}^{T} P(w(t)|q(t))P(q(t)|q(t-1)). \tag{3.6}$$

through training and model selection.

LSA-HMM method

LSA-HMM is based on solving the optimization task (3.6) by using δ parametrization for the *transition probabilities* $P(q(t)|q(t-1))$ and applying *latent semantic analysis* (Deerwester et al., 1990) (LSA) for the *emission probabilities* $P(w(t)|q(t))$ (Figure 3.8). Here we aim to obtain these conditional probabilities in a minimally-supervised manner, which does not require annotated training data. LSA-HMM is rather minimally supervised

than fully unsupervised, because it contains certain parameterizations and retains the capability for user-specified topics of interest. Further, the probability assumptions in HMMs are relaxed in LSA-HMM. To simplify the notation, we will refer in the following text, whenever possible, to the conditional probabilities $P(w_j|q_i)$ and $P(q_j|q_i)$ without the sequence index t.

Transition probabilities are based on a *self-transition probability* parameter $\delta \in (0, 1)$ that controls the segmentation granularity. The transition probability is then defined as

$$P(q_j|q_i) = \begin{cases} \delta & \text{if } j = i, \\ \frac{1-\delta}{N_q-1} & \text{if } j \neq i. \end{cases}$$

The probability of continuing the current topic is thus δ, and the remaining probability $1 - \delta$ of switching a topic is distributed uniformly to avoid supervised modeling assumptions.

Emission probabilities are based on LSA. LSA is a commonly applied technique for inducing text similarity measures from co-occurrence statistics in a large, unannotated data set. We use an LSA-based term-term similarity measure. Because the topic keywords occur in the majority of shift-wise documents and, more importantly, because different topics tend to co-occur in a single document, we apply the Word Space model (Schütze, 1998). It generates a term-by-term matrix and only considers word co-occurrence within a fixed context window. This allows sub-document distributional properties to be accounted for.

We derive the value of the emission probability $P(w_j|q_i)$ from the LSA similarity $\text{lsa}(w_j, q_i)$ of the word w_j to the topic q_i. This is based on the intuition of, for example, $P(pulse|hemodynamics) >> P(sister|hemodynamics)$.

We re-scale the LSA values in order to improve numerical comparability across topics; the original LSA values are mutually incomparable. For example, the top terms and their original LSA values for the topic of *Hemodynamics* are

1. *hemodynamiikka* [*hemodynamics*] (1.000),
2. *pulssi* [*pulse*] (0.910),
3. *sr* [*sr*, sinus rhythm] (0.819),
4. *rr-taso* [*rr-level*, respiratory rate] (0.785),
5. *korkeahko* [*quite high*] (0.784),
6. *sinusrythmi* [*sinus rhythm*] (0.784),
7. *rr* [*rr*, ambiguous abbreviation (e.g., respiratory rate, regular respirations, regular rhythm, Riva-Rocci, relative response, or relative risk] (0.768),

8. *verenpaine* [*blood pressure*] (0.716),
9. *lisälyönti* [*extrasystole*] (0.673), and
10. *ok* [*ok*] (0.672).

Here, the similarity of the terms *korkeahko* [*quite high*] and *sinusrythmi* [*sinus rhythm*] to the topic of *Hemodynamics* is indeed the same. However, comparison across the topics is meaningless: For example, for the topic of *Other*, defined as *other NOT breathing NOT ... NOT diuresis*, the top term *vatsa* [*stomach*] has the original LSA value of 0.683. The re-scaling is based on shifting the LSA value of the top term to 1.000, specifying the value where the maximum LSA similarity to another topic is larger than for a given topic, and assigning a minimal similarity of any word to any topic (Figure 3.9). See Ginter et al. (2008) for a more detailed description of re-scaling.

The final LSA-HMM model combines transition and emission probabilities as defined above. This preserves the overall structure of HMM but replaces the emission probabilities with a quantity that is not a probability. The optimal state sequence is obtained by solving

$$\arg\max_{q\in\mathcal{Q}} \mathrm{lsa}(w(1), q(1)) \prod_{t=2}^{T} \mathrm{lsa}(w(t), q(t)) P(q(t)|q(t-1)).$$

3.2.8 Performance evaluation

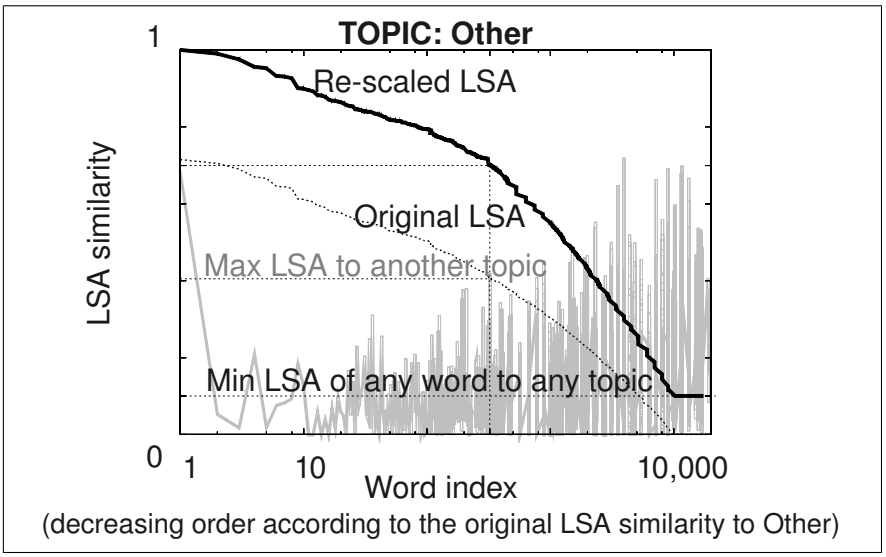
The performance is evaluated by comparing three methods in terms of the micro-averaged accuracy. The experiment is repeated with various amounts of training data for HMM. Also the macro-averaged WindowDiff is reported, but it evaluates topic segmentation quality independently of the topic labels. The WindowDiff window size was set to half of the average segment size in the manually annotated data, which is a standard way to set this parameter.

LSA is performed with the Infomap NLP software[4] (Dorow and Widdows, 2003) on all text available in the 448 patient reports from which no shift was selected for testing. The parameters for which supervision is needed are δ, the left and right LSA context window width, and LSA re-scaling parameters. Their values are specified by a grid search on 60 annotated shifts. To avoid over-fitting, the parameter selection shifts are held out from testing.

3.2.9 Evaluation results

The accuracy of the LSA-HMM was considerably better than that of the keyword search, but, as expected, it was outperformed by HMM (maximal

[4]See `http://infomap-nlp.sourceforge.net/` [cited 2009 August 1].



Adapted from Paper IV

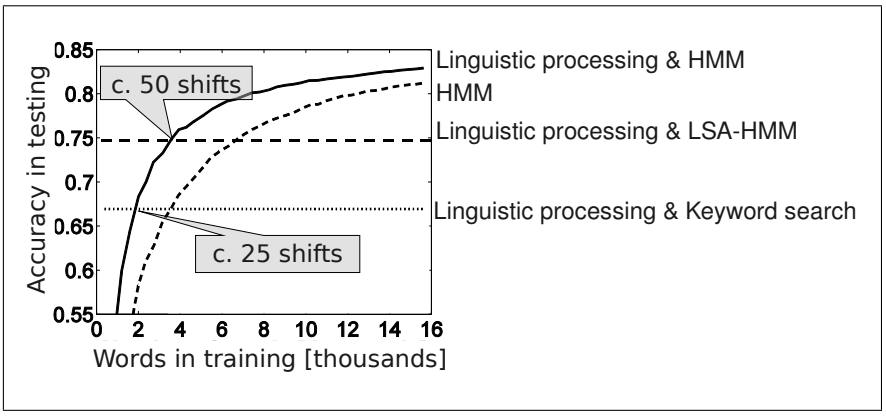
Figure 3.9: *Re-scaling the LSA values*

accuracy on the test set of 204 shifts (c. 16,000 words) 0.67, 0.75, and 0.83 for the keyword search, LSA-HMM, and HMM, respectively (Figure 3.10)). The supervised HMM method was allowed to learn from manually performed "model solutions", as opposed to LSA-HMM, which received only one keyword per topic.

To reach the performance of LSA-HMM, HMM required approximately $3,600$ words (c. 50 shifts) of manually labeled training data.

While not intuitive, the WindowDiff results (maximal values on the test set of 204 shifts 0.16, 0.23, and 0.21 for the keyword search, LSA-HMM, and HMM, respectively) were in disagreement with the accuracy results. The keyword search resulted in better WindowDiff performance than even HMM, and the performance of LSA-HMM and HMM was approximately the same. However, WindowDiff does not take into account the assigned labels, our key evaluation criterion. Thus, we do not view the WindowDiff results as compromising the positive primary findings in terms of accuracy.

Linguistic processing improved the performance (Figure 3.10). However, as expected, its significance diminished with increasing amount of training data.



Adapted from Paper IV

Figure 3.10: *Comparison of HMM and LSA-HMM methods and the benefit of linguistic processing*

3.2.10 Summary

The section has described applications of HMMs to topic segmentation and labeling of patient records. It has addressed the clinical needs by considering topics that are crucial for creating overviews, developing a minimally supervised method, evaluating the amount of manual annotation work, and assessing the importance of domain-tailoring. The proposed minimally supervised method is applicable to information search tasks with freely-chosen topics and very little labeled data available. However, if the search topics are established and resources for manual labeling exist, the supervised method should be preferred because it offers better performance. From the method development perspective, the results illustrate the differences between the topics. For example, the segment length and the number of most relevant terms vary.

3.3 Diagnosis coding

This section describes a multi-label classification application for the automated assignment of diagnostic codes to radiology reports. The application is related to the *Profile building*, *Attention focusing* and *Summarization* components of the road map for developing HLT (Figure 2.6) and the phase of using narratives (Figure 2.7). Sections 3.3.1, 3.3.2, and 3.1.3 describe the clinical application, learning task, and task-specific performance evaluation measures. In Section 3.3.4 the related work is discussed. Sections 3.3.5, 3.3.6, 3.3.7, and 3.3.8 specify the data consisting of US radiology reports,

linguistic processing, learning methods, and performance evaluation setting, respectively. Finally, Sections 3.3.9 and 3.3.10 report and summarize the results. Section 3.3 is based on Paper V.

3.3.1 Clinical application

The task is to assign the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (National Center for Health Statistics, 2007) codes to English free-text radiology reports automatically (Figure 3.11). All applicable codes are to be assigned to each report.

The task is given in the international medical natural language processing challenge (Computational Medicine Center, 2007), and it is motivated by US hospital administration and health insurance services; the assigned codes serve as justification for having specific procedures performed. But the application has also other uses: Both assigning codes and verifying previously given codes is possible. The first application can make administration more efficient, if the codes have not been entered to the patient record during the care process. The second application can be used in direct care to verify the given diagnoses and generate alarms for potentially missing or fallacious codes.

3.3.2 Learning task

In *multi-label classification*, as opposed to binary classification, multiple labels (e.g., *Breathing*, *Blood circulation*, and *Pain*) are considered. In binary classification there was one topic-label (e.g., *Breathing*) and the task was to assign instances either to the class of topically related (i.e., belongs to the class of *Breathing*) or to the class of topically unrelated (i.e., does not belong to the class of *Breathing*) objects. In other words, the multi-label classification task can be restructured to multiple binary classification tasks — one for each topic-label.

The difference in the multi-class classification is that in multi-label classification, each instance can belong to several classes at the same time. For example, in the multi-class classification task considered in Section 3.2, the constraint was that only one topic-label can be assigned to each text segment. In contrast, the topic labeling task in Section 3.1 belongs to the multi-label classification family, because every segment could have been relevant for all three topics.

a)
CLINICAL HISTORY
Eleven year old with ALL, bone marrow transplant on Jan. 2, now with three day history of cough.
IMPRESSION
1. No focal pneumonia. Likely chronic changes at the left lung base. 2. Mild anterior wedging of the thoracic vertebral bodies.
ICD-9-CM CODING
786.2 Cough

b)
CLINICAL HISTORY
This is a 7-month - old male with wheezing.
IMPRESSION
Borderline hyperinflation with left lower lobe atelectasis versus pneumonia. Clinical correlation would be helpful. Unless there is clinical information supporting pneumonia such as fever and cough, I favor atelectasis.
ICD-9-CM CODING
486 Pneumonia, organism unspecified
518.0 Pulmonary collapse
786.07 Wheezing

c)
CLINICAL HISTORY
7-year - old with history of reflux and multiple urinary tract infections.
IMPRESSION
Interval growth of normal appearing kidneys.
ICD-9-CM CODING
V13.02 Personal history, urinary (tract) infection

d)
CLINICAL HISTORY
One UTI. Siblings with reflux.
IMPRESSION
Normal renal ultrasound.
ICD-9-CM CODING
599.0 Urinary tract infection, site not specified

Reprint from Paper V

Figure 3.11: *Diagnosis coding task: narratives and codes*

3.3.3 Performance evaluation measures

In the challenge task, the aim is to evaluate the overall application quality, and not to consider performance separately for individual binary diagnosis coding tasks. The classification measures discussed in Section 3.1.3 can be adjusted for this multi-label classification setting, and then macro- or micro-averaged, as described in Section 3.2.3.

In addition to these measures, *cost-sensitive accuracy* is used in the challenge. The organizers motivate this by clinical regulations enforcing over and under-coding penalties to avoid additional risks of possible prosecution for fraud and lost revenues (Pestian et al., 2007). If B_i is the number of instances that are assigned to the topic i either in $f(X)$ or Y, then cost-sensitive accuracy

$$CSA = \left(1 - \frac{p_u FN' + p_o FP'}{\sum_{i=1}^{n_c} B_i}\right)^c .$$

In the challenge the constant $c = 1$, over-coding penalty $p_o = 1$, and under-coding penalty $p_u = 0.33$.

3.3.4 Related work

Automated diagnosis coding has attracted wide attention both among health care practitioners and academic researchers, especially in the USA. Examples of particularly successful automated diagnosis coding applications are MedLEE, Medical Language Extraction and Encoding System and Autocoder: MedLEE is routinely used in the New York Presbyterian Hospital to parse English patient records and map them to Unified Medical Language System (Bodenreider, 2004) (UMLS) codes (Mendonça et al., 2005). Adapting it for ICD-9-CM coding has also been studied (Lussier et al., 2000). Autocoder is implemented at the Mayo Clinic in Rochester, Minnesota to assign unit specific ICD-9-CM-related codes to patient records and it has resulted in a change in the coding personnel's duties to code verification and an 80% workload reduction (Pakhomov et al., 2006). In addition, for example, Pakhomov et al. (2006); Hripcsak et al. (2007); Pakhomov et al. (2007b) and Pakhomov et al. (2007a) discussed in Section 3.1.4 study automated recognition of patients with a given diagnosis.

Also the challenge has been popular. According to the report of the organizers (Pestian et al., 2007), the number of registrations is approximately 150, from six continents and more than twenty countries. The number of final submissions is 44. The organizers' review of all submissions suggests that for this particular task, the choice of the classifier is not crucial for success, but linguistic processing plays a key role; use of negations, the structure of UMLS, hypernyms, synonyms, and symbolic processing contributes to the

performance. More information about highly ranked submissions can be found, for example, in Farkas and Szarvas (2008) (the first place), Goldstein et al. (2007) (the second place), Paper VI (the third place), and Crammer et al. (2007) (the fourth place).

3.3.5 Patient record data

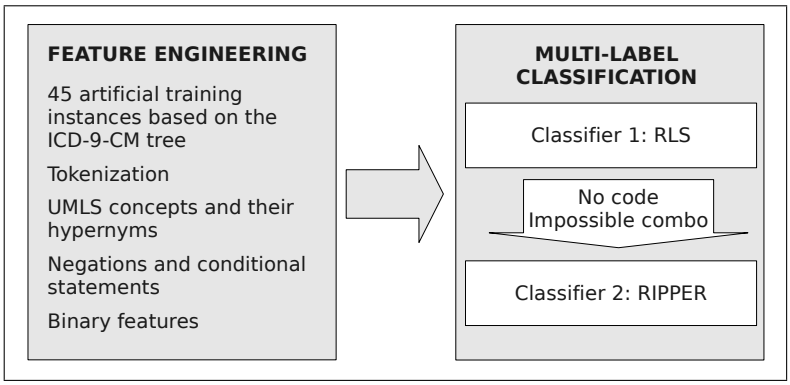
The anonymized patient record data (Computational Medicine Center, 2007) is collected from a US radiology department for children. The data has $1,954$ patient records divided into the training set of 978 records and the test set of 976. They are written in English and describe chest x-ray and renal procedures. The data consists of two subsections of the original patient records that are seen as fundamental for assigning the ICD-9-CM codes: clinical history provided by an ordering physician before the procedure and impression reported by a radiologist after the procedure (Figure 3.11). Similarly to the ICU narratives, the style of these radiology reports is concise and highly specific.

The data is accompanied with gold standard ICD-9-CM code annotation obtained by a majority vote of three independent parties. The majority vote is selected because the nature of the coding task is ambiguous: even though official coding guidelines (Moisio, 2000, pp. 69–126) exist, unit-specific detailed instructions are used to complement them. For example, the official guidelines state that uncertain codes should not be assigned, a definite diagnosis should be specified, when possible, and symptoms must not be coded if the definite diagnosis is available.

Altogether 45 different codes in 94 combinations are present in the data. The most common are

1. 786.2 *Cough* ($n = 310$),
2. 599.0 *Urinary tract infection, site not specified* ($n = 193$),
3. 593.70 *Vesicoureteral reflux, unspecified or without reflux nephropathy* ($n = 161$),
4. 780.6 *Fever* AND 786.2 *Cough* ($n = 151$), and
5. 486 *Pneumonia, organism unspecified* ($n = 132$).

The test set is released without ICD-9-CM coding, but it is restricted by requiring that any combination of codes occurs at least once both in the training and test data.



Adapted from Paper V

Figure 3.12: *Flow chart of the components*

3.3.6 Linguistic processing

Our application for automated assignment of ICD-9-CM codes to free-text radiology reports can be divided into two phases (Figure 3.12): This section describes the feature engineering phase, where text is enriched through linguistic processing and features that improve performance are extracted from the input text. Section 3.3.7 then describes the classification phase, where a cascade of two classifiers is used to assign the ICD-9-CM codes.

To start the linguistic processing, the training set is augmented with a small set of artificial instances. Then, text is tokenized and the patient records are represented as bag-of-words vectors. Then all data is semantically enriched using UMLS concepts and their hypernyms. Further, features are marked for occurrence in a negative or conditional context. Finally, the documents are represented as a set of binary features. This feature engineering is explained in more detail next.

The training set is augmented with 45 artificial instances obtained by concatenating the textual description of each of the 45 codes used in the challenge with the descriptions of its parents in the ICD-9-CM tree. For example, the artificial instance corresponding to the code 593.70 *Vesicoureteral reflux, unspecified or without reflux nephropathy* is *Diseases Of The Genitourinary System. Other diseases of urinary system. Other disorders of kidney and ureter. Vesicoureteral reflux. Vesicoureteral reflux unspecified or without reflux nephropathy.* The strategy is based on the observation of informative keywords appearing both in the radiology reports corresponding to a given code and in the description of this code in the ICD-9-CM tree (see, e.g., Figure 3.11a and 3.11c).

The text is tokenized and UMLS concepts occurring in the text are recognized using the MetaMap program (Bodenreider, 2004). This reduces the data sparseness, because, for example, related expressions *pneumonia* and *superimposed pneumonia* are both represented by the UMLS concept *C0032285*. In addition, hypernyms of the UMLS concepts are also used to enrich the data. Thus, the pneumonia-related UMLS concept *C0032285* is augmented so that it also contains the concept codes for *respiratory tract infection*, *disease caused by microorganism*, *bacterial infection*, and other hypernyms.

Contexts signaling uncertain or negative findings are identified in the text using a list of common trigger expressions such as *no*, *possible*, *suggestive*, and *likely*. In accordance with the coding guidelines, the goal is to avoid assigning negative or uncertain codes. The scope of the identified negation or conditional statement is assumed to continue to the end of the sentence. Hypernyms are processed with special care so that, for example, *no pneumonia* does not imply *no respiratory tract infection*, as other respiratory tract infections may be present.

3.3.7 Learning methods

A machine-learning approach using a cascade of two classifiers trained on the same data is used to predict the codes. Both classifiers are trained with the same data, and perform multi-label classification by decomposing the task into 45 binary classification problems, one for each code.

In this setting, it is possible for a classifier to predict an empty, or impossible, combination of codes. Such known errors are used to trigger the cascade: when the first classifier makes a known error, the output of the second classifier is used instead as the final prediction. No further correction of the output of the second classifier is performed as preliminary experiments suggested that this would not further improve the performance.

The first classification method is RLS (see Section 3.1.7). RLS is used in the challenge application because it has the following computational advantages (Pahikkala, 2008): Firstly, it is possible to calculate the cross-validation performance of RLS on the training data without retraining in each repetition. Secondly, the RLS solution can be computed for several different values of the regularization parameter as efficiently as calculating for only one. Thirdly, several learning problems on the same data set can be solved in parallel, provided that the same kernel function is used with each problem, as is the case in the challenge task. These properties are helpful in testing which strategies improve the performance.

The second method in the cascade is the *RIPPER* rule induction-based learning method (Cohen, 1995). The *rules* learned by the algorithm are formulated in propositional logic. Each individual rule is a conjunction of

individual *conditions*, which can be of the form $A = v$ (A being a nominal attribute), or $A \leq v$ or $A \geq v$ (A being a real valued attribute), where A is a nominal attribute, and v a value. Rules are sequences learned for recognizing positive instances, and they are applied in the class prediction. RIPPER is used in the challenge application due to its excellent performance with the challenge data and because it and RLS have quite differing learning principles. Consequently, RIPPER may succeed in cases where RLS fails. RIPPER was not, however, chosen as primary classifier, because RLS performs slightly better.

3.3.8 Performance evaluation

The primary performance measure is micro-averaged $F1$ ($F1_{\text{micro}}$), which is the official challenge measure. In addition, the organizers have reported macro-averaged $F1$ ($F1_{\text{macro}}$) and *CSA*, but they have no effect on the submission ranking.

The final evaluation is performed by the organizer through a comparison of the system output and the majority vote gold standard on the test set 976 patient records. In the method development, different strategies to improve the performance is tested by evaluating their effects using a 10-fold cross-validation on the training set of 978 patient records.

3.3.9 Evaluation results

The learning task was difficult as illustrated by the small number of final submissions (44 final submissions vs. approximately 150 registrations). The mean, standard deviation and median of the 44 final submissions in terms of $F1_{\text{micro}}$ in the test set were 0.767, 0.133 and 0.799, respectively (Pestian et al., 2007). By comparison, the pairwise inter-annotator agreement also measured by using $F1_{\text{micro}}$ in the test set was between 0.826 and 0.896, when comparing individual annotators' opinions against the gold standard (note that because the gold standard was formed based on these annotations, the numbers are likely to give an over-optimistic view on agreement between the gold standard and content experts' opinions on the coding task, in general), and between 0.673 and 0.758 in the pairwise comparisons of the individual annotators' opinions (Farkas and Szarvas, 2008). This advocates using machine learning to assign ICD-9-CM codes in clinical practice; the performance of the top ten submissions was between 0.850 and 0.891 (Table 3.6).

Our application scored, with $F1_{\text{micro}} = 0.877$, the third place in the challenge. While developing the application, the focus was on maximizing $F1_{\text{micro}}$, the official challenge measure. However, different performance evaluation measures emphasize different aspects: $F1_{\text{micro}} = 0.877$ illustrates a good quality in relation to the number of code assignments. In terms of

Table 3.6: *Performance of top ten challenge submissions trained with* 978 *patient records and tested with* 976 *records*

Adapted from Computational Medicine Center (2007)

Rank	$F1_{\mathbf{micro}}$	$F1_{\mathbf{macro}}$	**Cost-sensitive accuracy**
1	0.891	0.769	0.918
2	0.886	0.729	0.909
3 (our submission)	0.877	0.703	0.913
4	0.876	0.721	0.909
5	0.872	0.776	0.901
6	0.871	0.733	0.898
7	0.868	0.732	0.900
8	0.859	0.668	0.905
9	0.851	0.682	0.901
10	0.850	0.676	0.878

the over and under-coding penalty-bearing *CSA* measure, the performance was even better, when compared to the other submissions. $F1_{\text{macro}} = 0.703$ would have given only the seventh place, but this measure evaluates the performance in all 45 classes, and maximizing this was not the aim in the challenge. According to the report of the challenge organizers (Pestian et al., 2007), the choice of the classifier was not crucial for success, but use of negations, the structure of UMLS, hypernyms, synonyms, and symbolic processing seemed to contribute to performance. These characteristics were also evident in our method development.

When developing the application, a modular approach was adopted (Table 3.7); different strategies to improve the performance were tested trough cascading them. Only the modules that led to a better performance were included. The good performance of the conjunctive rules (i.e., the RIPPER module) implied that the coding task could be reduced to recognizing certain groups of informative keywords from the text (see also Farkas and Szarvas (2008)). This justified the module of artificial ICD-9-CM instances. Regarding the cascade of two classifiers, the identification of impossible code combinations might not be as straightforward in a real-world setting as it was in the challenge. The cascade is still likely to be applicable in a majority of cases: Approximately 50 % of known errors triggering the cascade were due to RLS not giving a code. In addition, the ICD-9-CM tree could be used as a trigger, because, for example, codes both to a disease and its symptom should not be assigned.

Table 3.7: *Estimate of the effect of the cascaded modules on overall performance in the ten-fold cross-validation on the training set of 978 patient records*

Reprint from Paper V

Module	$1 - F1_{\mathbf{micro}}$	**Relative decrease against the previous row**
RLS	0.207	-
Tokenization	0.193	0.07
UMLS concepts	0.175	0.09
UMLS hypernyms	0.166	0.05
Negations and conditional statements	0.153	0.08
RIPPER	0.135	0.12
ICD-9-CM instances	0.134	0.01

3.3.10 Summary

The section has introduced a multi-label classification application for the automated assignment of diagnostic codes to radiology reports. The application demonstrates the usefulness of machine learning for clinical practice and the importance of domain-tailoring. The experiences gained from the challenge are beneficial for developing tools for real-world use. In addition, the section concretizes multi-label classification tasks and the related performance evaluation.

Chapter 4

Performance Evaluation Methods and Evaluation Reliability

This chapter addresses performance evaluation methods and evaluation reliability. It is combines Papers VI and VII by summarizing the main results.

Paper VI discusses ways to ensure evaluation reliability in assessing learning performance. It clarifies the steps of the performance evaluation process (Figure 1.3) and their associations in classification applications. In particular, the study addresses the *AUC* measure and cross-validation method. The paper has also been the motivation for Paper VII and reference for Section 4.3.3.

Paper VII considers the problem of evaluating performance reliably with the constraints of having a regression task, an overwhelming amount of data, limited processing time, a supervised setting, and the RLS algorithm. It addresses this problem by adopting a faster learning algorithm (i.e., the sparse version of the RLS algorithm), developing a faster performance evaluation method (i.e., an efficient hold-out method) for this algorithm, and showing that this method contributes not only to processing time but also to the evaluation diversity and quality.

See Paper VI for a more elaborate discussion on all steps of the performance evaluation process. Paper VII proves the algorithms and computational complexities that are presented next.

4.1 Sparse regularized least-squares algorithm

As explained in Section 3.1.7, the RLS algorithm has been shown to have a state-of-the-art performance in regression and classification, and it is often used in practical applications. However, the computational complexity

of training an RLS learner, $O(m^3)$, may be too tedious, if the number of training instances, m, is large. To address this problem, several authors have considered sparse versions of RLS in which only a part of the training instances, often called the *basis vectors*, are used as regressors while the whole set is still used in the training process (see, e.g., Smola and Schölkopf (2000) and Rifkin (2002)). This decreases the training complexity of RLS to $O(mn^2)$, where $n << m$ is the number of basis vectors. Next, to serve as an introduction to our subsequent original contribution, an existing solution to improve the complexity is described.

The proof of the decreased training complexity can be summarized as follows: Let us first describe (3.5) and (3.4) in a matrix form. When $X = (x_1, \ldots, x_m)$,

$$\begin{aligned} f(x) &= \sum_{i=1}^{m} \alpha_i k(x, x_i) = (k(x, x_1), \ldots, k(x, x_m))(\alpha_1, \ldots, \alpha_m)^{\mathrm{T}} \\ &= k(x, X)A, \\ f_A(X) &= ((k(x_1, X)A, \ldots, k(x_m, X)A)^{\mathrm{T}} = KA, \text{ and} \\ \|f\|_k &= A^{\mathrm{T}}KA. \end{aligned}$$

Then the RLS minimization task (3.4) is

$$\arg\min_A g(A) = \arg\min_A (Y - KA)^{\mathrm{T}}(Y - KA) + \lambda A^{\mathrm{T}}KA,$$

where $Y = (y_1, \ldots, y_m)$. This is solved by setting the derivative $dg(A)/dA$ to zero. The solution is

$$A = (KK + \lambda K)^{-1}KY$$

whose computational complexity is $O(m^3)$ because of inverting the $m \times m$ matrix $(KK + \lambda K)$.

In the sparse version, only the basis vectors $B \in 1, \ldots, m$ have $\alpha_i \neq 0$ in (3.5). When $|B| = n$, M_R denote a sub-matrix of M with rows indexed by R, and M_{RC} a sub-matrix of M with rows indexed by R and columns indexed by C, this results in the *sparse RLS* minimization task

$$\arg\min_A g(A) = \arg\min_A (Y - (K_B)^{\mathrm{T}}A)^{\mathrm{T}}(Y - (K_B)^{\mathrm{T}}A) + \lambda A^{\mathrm{T}}K_{BB}A,$$

because

$$\begin{aligned} f(x) &= \sum_{i \in B} \alpha_i k(x, x_i), \\ f_A(X) &= (K_B)^{\mathrm{T}}A, \text{ and} \\ \|f\|_k &= A^{\mathrm{T}}K_{BB}A. \end{aligned}$$

Again, the solution is obtained by setting the derivative $dg(A)/dA$ to zero. The solution is

$$A = (K_B(K_B)^{\mathrm{T}} + \lambda K_{BB})^{-1} K_B Y.$$

Now, the computational complexity is dominated by computing $K_B(K_B)^{\mathrm{T}}$: the complexity of $K_B(K_B)^{\mathrm{T}}$ is $O(mn^2)$, whereas the inversion

$$(K_B(K_B)^{\mathrm{T}} + \lambda K_{BB})^{-1}$$

can be performed in $O(n^3)$ time with $n << m$.

4.2 Efficient hold-out method

A complementary approach to developing faster learning algorithms is to address computational efficiency of performance evaluation methods used in model selection and testing. Hold-out techniques, in particular cross-validation, are among the most commonly used of these. Recently, fast cross-validation algorithms have been proposed for many learning algorithms (see, e.g., Mullin and Sukthankar (2000) and Pahikkala et al. (2009)).

In Paper VII, we develop en efficient cross-validation method for the sparse RLS. Let $F = \{1, \ldots, m\}$ index the whole data, $B \subseteq F$ denote the basis vectors, $H \subseteq F$ denote the hold-out set, $E = H \cap B$ the intersection of the hold-out set and the basis vectors, and $L = \overline{H} \cap B$ the intersection of the non-hold-out set and the basis vectors. Then predictions for the hold-out set can be calculated from

$$f_{\overline{H}}(X_H) = K_{HL}(K_{L\overline{H}} K_{\overline{H}L} + \lambda K_{LL})^{-1} K_{\overline{H}L} Y_{\overline{H}}.$$

This can be implemented efficiently given that Sparse RLS is trained in advance for the whole data F in $O(mn^2)$ time: the computational complexity of holding out the set H is $O(|H|^3 + |H|^2 n)$, and those of N-fold cross-validation and leave-one-out cross-validation are $O(m^3/N^2 + m^2 n/N)$ and $O(mn)$, respectively.

When compared with the previously proposed cross-validation method for the sparse RLS (Cawley and Talbot, 2004), our method has several advantages. First, our method is a hold-out method, which allows holding out several data points simultaneously. Hence, it can be applied to cross-validation in general, unlike the Cawley and Talbot's leave-one-out cross-validation method.

The second difference is related to the computational efficiency: In leave-one-out cross-validation, the computational complexity of Cawley and Talbot's method is $O(mn^2)$, while that of ours is only $O(mn)$. Further, our algorithm can be combined with the simultaneous training of the sparse RLS with several values of λ so that this model selection can be performed as efficiently as training just one instance of the sparse RLS.

Third, our method allows holding out basis vectors too. Namely, Cawley and Talbot's method seems to assume that basis vectors are chosen first from the whole training set and then leave-one-out cross-validation is performed by holding out each of the remaining data points in turn. However, in certain situations, a need for holding out basis vectors arises and consequently, we assume that basis vectors and hold-out data can be selected independently. Our method can also be used, when a basis vector is chosen to be held out. The need for holding out basis vectors is elaborated next by addressing cross-validation more thoroughly.

4.3 Evaluation reliability

This section motivates the use of our hold-out method from the reliability point of view. First, it discusses different aims for cross-validation and then three different cross-validation techniques.

4.3.1 Aims for cross-validation

Cross-validation can be used to find answers for various statistical questions in machine learning. The question at hand determines the way cross-validation should be used. Dietterich (1998) represents a taxonomy of the questions. In this taxonomy, the first dividing factor is whether the machine learning problem in question considers a single application domain or multiple domains. Our evaluation addresses machine learning in a single application domain, as opposed to using the same algorithm for various learning tasks such as classification related to finance and health care. The next division in the single domain branch of the taxonomy is between evaluating

(a) how good are the predictors the learning algorithm usually creates, or

(b) how good is the predictor learned with the algorithm from a certain training set.

In other words, the question (a) addresses the quality of a certain algorithm in a given domain is evaluated whilst the question (b) corresponds to evaluating a certain trained predictor.

For purposes of question (a), ten times repeated ten-fold cross-validation is recommended (Kohavi, 1995). Evidence supporting the selection of a repeated cross-validation technique is also given in (Dietterich, 1998). The justification is that the aim is to measure the quality of a learning algorithm with unspecified user needs, and hence the evaluation method should take into account the variability due to the choice of training and test sets.

In contrast, for purposes of the question (b) is considered, the training set variability should be eliminated. This is because the predictor to be

evaluated has been already trained and hence, also the user needs are known. Leave-one-out cross-validation is particularly good for this purpose. It is an almost unbiased estimator of the prediction performance, that is, the final evaluation measure value is biased due to the removal of only one instance from the training set.

4.3.2 N-fold and leave-one-out cross-validation

Let us now consider the evaluation task of answering to the question (a) with the sparse RLS by using N-fold cross-validation. For N-fold cross-validation, we prefer random selection of basis vectors, because using a selection method that takes advantage of the whole training set outside cross-validation repetitions may cause biased performance estimators.

In our evaluation task, the variability caused by the selection of the basis vectors should be taken into account in addition to the variability caused by the training and test sets. This is emphasized with randomly selected basis vectors. Taking the variability caused by the basis vector set into account can be accomplished, for example, by selecting the basis vectors randomly and separately for each cross-validation repetition. If we want, at the same time, to preserve the computational efficiency, we can take into account the variability caused by the selection of the basis vectors to some extent, if we hold out part of the basis vectors in each cross-validation repetition.

For example, let us assume the evaluation task of assessing the sparse RLS having n randomly selected basis vectors with the ten-fold cross-validation. We can start by training the sparse RLS with $11n/10$ randomly selected basis vectors. Then, we hold out $n/10$ of the basis vectors and one tenth of the non-basis vectors in each repetition. Consequently, this cross-validation provides an approximation of the standard cross-validation estimator in which one tenth of the basis vectors is changed in each repetition instead of the whole set of basis vectors being changed. Compared to selecting all basis vectors completely randomly for each cross-validation repetition, this approach retains the computational efficiency, because we have no need to train the sparse RLS again with new basis vectors.

Let us now consider the evaluation task of answering to the question (b) with the sparse RLS trained with a fixed data set and basis vectors. Now, the same set of basis vectors should be used in each cross-validation repetition. Therefore, it makes sense to hold out only non-basis vectors and have hold-out sets as small as possible, the extreme case being leave-one-out cross-validation. Naturally, the fastest leave-one-out cross-validation method results in the fastest training, model selection, and testing process.

4.3.3 Leave-cluster-out cross-validation

For reliable performance evaluation, the test set has to be completely independent from the creation of the learner. With hold-out techniques, this means assuring that no information leak between training, validation and test sets exists. However, it is not easy in practice to notice all task-specific data dependences that have to be taken into account when holding out data.

To illustrate the difficulties, let us consider the task of clinical topic labeling. At least four types of task-specific semantic dependences must be taken into account when holding out data: *patient profile*, *author*, *community*, and *time*. The aim is to label according to the topic-segment similarity. Hence, the dividing factor should not be the patient profile, the author, the community in which the text was written nor its writing time.

If the aim is to build a topic classification application for documents of several patients, the documents of the same patient should be used only for training, validation, or testing: A potential information leak occurs because narratives about one patient are likely to be very homogeneous which makes it easier for a learner to recognize notes relevant to a given topic if the test set contain data about same patients as the training and validation sets. This principle of independent training, validation, and test sets was followed in Chapter 3.

Each author has an individual writing style, and as a result, texts written by the same person are very likely to be more similar than those written by two different people. If the aim is to build a topic classification application for documents written by many nurses, the notes written by the same nurse should be used only for training, validation, or testing. This was not possible in Chapter 3 because the data did not include information regarding the author.

Similarly, texts reflect the surrounding community and time, when it was written. If the aim is to build a topic classification application for documents of several health care units, the notes written in the same unit should be used only for training, validation, or testing. Finally, the guideline for time-dependence is to test the performance regularly and re-train the classifier if necessary.

These examples illustrate that the assumption often made in the machine learning studies of having independently and identically distributed data is not met in many practical tasks. The conclusion is that for evaluation reliability, we need an evaluation method that takes the data dependences, or clusters, into account.

Leave-cluster-out cross-validation (Pahikkala et al., 2006) is a performance evaluation method that meets this need. Each fold in the leave-cluster-out cross-validation consists of the data points that form a cluster. That is, the method generalizes leave-one-out cross-validation by holding

out one cluster at a time. To explain the method and motivate its use for the sparse RLS to answer the question (b), let us consider four empirical examples.

Pahikkala et al. (2006) describes an experiment, where the effects of clustered data are illustrated by comparing the leave-one-out and leave-cluster-out cross-validation performance of a trained RLS. The task was selecting the best answer from a set of candidate answers, that is, ranking the candidates by the order of their correctness, in the case of sentence analysis. A training data set was obtained by generating a set of parse candidates for one hundred sentences. Each sentence had a known "correct" parse that a parser is supposed to output for the sentence. Each candidate parse was associated with a score value indicating its similarity to the correct parse. The task of the trained RLS was, for each of the one hundred sentences, to rank its candidate parses in the order of their score values. For this purpose, an RLS regressor was trained by using the generated parses and their score values.

In this example, the data set is heavily clustered according to the sentences the parses were generated from. Due to the feature representation of the parses, two parses originating from a same sentence almost always have larger mutual similarity than two parses originating from different sentences. Therefore, the clustered structure of the data has a strong effect on the performance evaluation results obtained by cross-validation: Data points that belong to the same cluster as the hold-out data point have a dominant effect on the prediction.

The problem can be solved by performing cross-validation on the sentence level so that all parses generated from a sentence are either always in the training set or in the test set, that is, by using the leave-cluster-out cross-validation. The experimental results confirm this conclusion.

Another example of taking clustered data into account at the method implementation step is discussed in Sætre et al. (2007). The study discusses the differences between two popular alternatives of doing ten-fold cross-validation by using data describing protein-protein interactions. The first alternative is to divide the data into ten groups before doing any analysis. The second alternative is to perform linguistic processing and feature extraction on the whole corpus and then divide data into cross-validation folds. The evidence in support of the serious information leak from creation of the learner to testing is shown in the second alternative: it gave substantially better impression of the performance.

As a third example of taking clustered data into account, we refer to Saeh et al. (2005), were the method of leave-cluster-out is used to improve evaluation reliability in classifying chemical compounds as active or inactive in biochemical processes of interest. A cluster is defined as active if it contains an active chemical compound. In leave-cluster-out, one active cluster at

a time is removed from training. Then, to further improve reliability, the authors apply *leave-core-out*, where compounds that contain the same core, that is, compounds that look alike, are left out one at a time. However, this study does not focus on comparing evaluation methods but to improve evaluation reliability in a particular machine learning application.

Finally, also Simon et al. (2003) emphasizes unreliability of performance evaluation if it contains steps not in line with cross-validation. Their experiments measure misclassifications in a genetic application when using two types of leave-one-out cross-validation alternatives. In the first alternative, hold-out instance is removed before any processing. In the second alternative, some analysis and processing is performed before the removal. As expected, the results underscore better reliability with the first approach.

4.4 Evaluation results

This section contains the results of empirical comparisons of methods to compute the cross-validation performance for the sparse RLS. In accordance with Section 4.3, N-fold and leave-cluster-out cross-validation methods are considered. To improve the reliability of the results, the experiments are performed using parallel implementations in the MATLAB and Python programming environments. In accordance with the sparse approach, large data sets with multiple features are used.

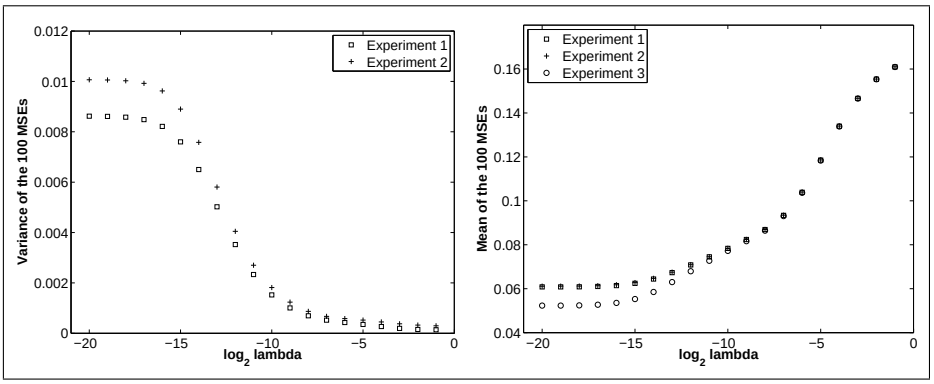
As a preliminary experiment, we have confirmed the difference between our and Cawley and Talbot's leave-one-out methods: The results differ only if holding out a basis vector. The experiment includes separate tests with the number of basis vectors $0.1m$, $0.3m$, and $0.5m$, and the difference between results is always present. The tasks in the experiment also comprise the empirical part of Cawley and Talbot (2004), but we do not consider them in the actual experiments, because larger data sets are better suited to statistical significance testing and should be used for sparse learning methods.

4.4.1 N-fold cross-validation

The variability caused by the basis vector selection was tested in four large-scale regression tasks: Ailerons, Elevators, Pole Telecomm and Pumadyn[1]. Ailerons contained $7,154$ instances and forty features per instance. The respective numbers were $8,752$ and eighteen for Elevators; $5,000$ and 48 for Pole Telecomm; and $4,499$ and 32 for Pumadyn. The kernel matrix K was formed by using a Gaussian radial basis function kernel

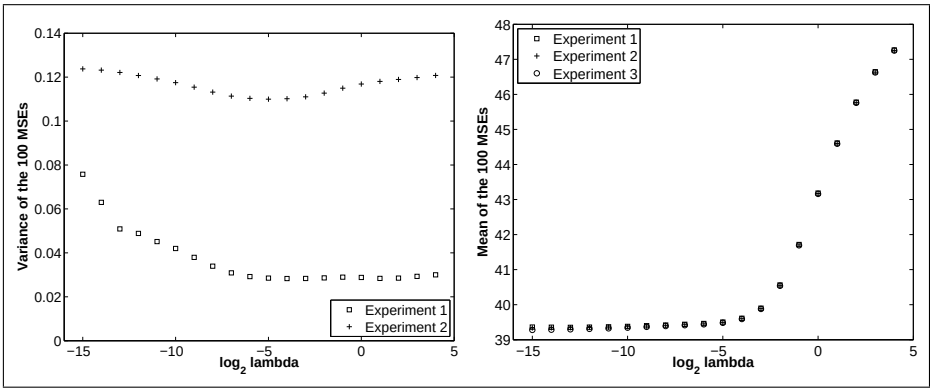
$$k(x, x') = \exp\left(-\gamma|x - x'|^2\right),$$

[1]Downloaded from `http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html` [cited 2009 February 23].



Reprint from Paper VII

Figure 4.1: *Variance and mean in the Ailerons task*



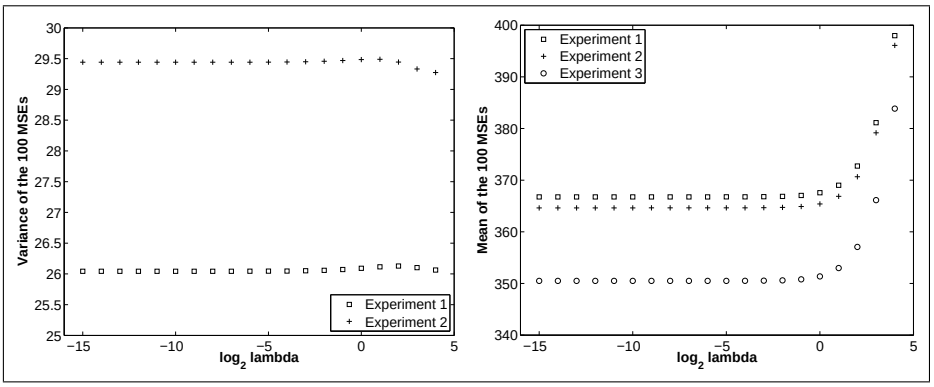
Reprint from Paper VII

Figure 4.2: *Variance and mean in the Elevators task*

where $\gamma \in \mathbb{R}$ is a positive constant determining the width of the kernel. Suitable γ values were selected for each task in a preliminary experiment. The positive definiteness assumption of K was assured by a diagonal shift of $10^{-7}I$, where I is an identity matrix of an appropriate order. The tested domain for λ was $\{2^{-20}, \ldots, 2^{-1}\}$ in the Ailerons task, and $\{2^{-15}, \ldots, 2^{4}\}$ in other three tasks.

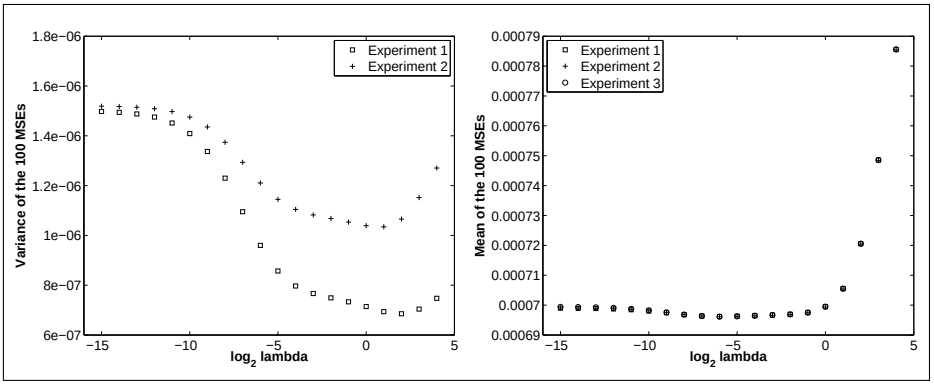
The analysis included three experiments:

1. In the first experiment, thirty basis vectors were selected randomly from the whole data. Then, the data was divided randomly into ten folds so that each fold contained three basis vectors. The sparse RLS regressor was trained and cross-validation error was computed using



Reprint from Paper VII

Figure 4.3: *Variance and mean in the Pole Telecomm task*



Reprint from Paper VII

Figure 4.4: *Variance and mean in the Pumadyn task*

the fold partition. This process was repeated one hundred times with different sets of basis vectors and different fold partitions.

2. In the second experiment, 27 basis vectors were selected randomly from the whole data. Then the data was divided randomly into ten folds so that no basis vectors were included in any fold. This experiment was also repeated one hundred times.

3. The third experiment was the same as the second one but with the thirty basis vectors selected initially in the first experiment.

The number of basis vectors was selected in a preliminary experiments in a way that it is suitable for solving the tasks.

The first two experiments were to confirm the hypothesis that a smaller variance for the cross-validation estimator is obtained by varying the set of basis vectors in each repetition. These experiments were used to evaluate the error that the sparse RLS regressor does, when it is trained with 27 basis vectors: In the first two experiments each cross-validation repetition contained 27 basis vectors. The sizes of the hold-out sets were approximately $m/10$ in the first experiment and $(m-27)/10$ in the second one, and the difference between these sizes is negligible with large m. Furthermore, three basis vectors were switched in each cross-validation repetition in the first experiments while the basis vectors were the same in each repetition in the second experiment. The effects caused by the selection of the basis vectors were tested by repetitions. The mean squared error was used as a performance evaluation measure, and the variance was computed with the following formula:

$$\sigma^2 = \frac{1}{1+r}\sum_{i=1}^{r}(MSE^{(i)} - \mu)^2, \tag{4.1}$$

where r is the number of repetitions, $MSE^{(i)}$ is *MSE* obtained from the ith repetition, and μ is the mean cross validation error estimated from the sample of repetitions. The results gave supportive evidence of obtaining a smaller variance by varying the set of basis vectors in each repetition (Figures 4.1–4.4). This was most evident in Elevator and Pumadyn tasks.

With the third experiment the benefit of having three additional basis vectors in training was addressed; the first two experiments had 27 basis vectors whilst the third one had thirty. The evaluation observed whether the differences between *MSE* values for the experiment pairs were statistically significant. When the first and third experiment were compared, the Wilcoxon signed-rank test was used, and for the other two pairs (i.e., 1st–2nd, 2nd–3rd) the Wilcoxon rank-sum test was chosen (Wilcoxon, 1945). These statistical tests were selected, because they do not assume a normal distribution. The rank-sum test was used, because the second experiment had a different set of basis vectors than the other two experiments.

In the results, additional basis vectors improved or did not change the learning performance. The differences between *MSE* values for the experiments with 27 basis vectors were statistically significant (i.e., two-tailed $p < 0.05$) only in six out of sixty comparisons. In contrast, the comparison of experiments with 27 and thirty basis vectors almost always produced statistically significant differences in Ailerons and Pole Telecomm tasks; three exceptions occurred in the Ailerons task with the largest and second largest λ values. In the Elevators task, the six smallest λ values produced statistically significant differences between the first and third experiment and between the second and third experiment. In the Pumadyn task, only the comparison between the first and third experiment with the five smallest λ

values had a statistically significant difference. As we expected, the extra basis vectors contributed to the performance (Figures 4.1–4.4).

4.4.2 Leave-cluster-out cross-validation

To study the effects of clustered data, the sentence analysis task of Section 4.3.3 is considered. The parse ranking data set[2] that had altogether $2,354$ instances was used: The total number of sentences was 501 and approximately five parse candidates were generated for each sentence. The feature representation contained tens of thousands of different features. The kernel matrix K was formed by using a linear kernel and ensured its positive definiteness by a diagonal shift of $10^{-7}I$. Due to the sentence-wise clustered data, cross-validation was performed on the sentence level by holding out one sentence at a time. One training instance per each sentence was chosen as a basis vector randomly, totaling $n = 501$. This selection was intuitive and its superiority over complete random selection was confirmed empirically in a preliminary experiment.

The *MSE* of the following two methods was compared:

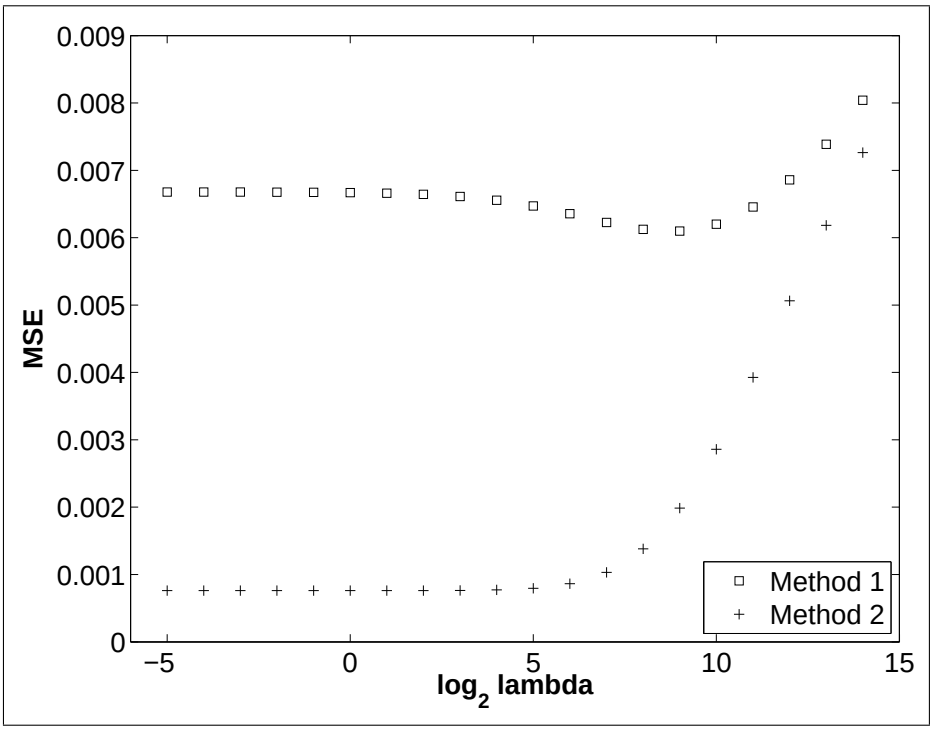
1. Our hold-out method, which was used so that the one cluster was completely removed from the training set in each cross-validation repetition. That is, the basis vectors were also removed.

2. A cross-validation method in which one cluster at a time was removed except the basis vector associated to the cluster, that is, the basis vector is preserved and the square loss is evaluated on it.

The Wilcoxon signed-rank test was used to evaluate whether the differences between *MSE* values for the two methods were statistically significant. Because the first method always held out one parse more than the second method (i.e., the basis vector), each sentence was analysed separately and the significance was calculated using the average *MSE* for the sentence. A separate comparison was made for each λ value from 2^{-5} to 2^{14}.

The experiments clearly indicated that the ability to hold-out basis vectors from training is necessary with clustered data (Figure 4.5). The second method always gave an over-optimistic *MSE*, because when a sentence was held out, the method was still allowed to learn from the respective basis vector. The difference in the *MSE* values was statistically significant for all twenty comparisons.

[2]Downloaded from `http://staff.cs.utu.fi/~aatapa/software/RLScore/` [cited 2009 February 23].



Reprint from Paper VII

Figure 4.5: *Cross-validation methods and the sentence analysis*

4.5 Summary

The chapter discussed various hold-out methods and introduced a new hold-out method and its implementation for the sparse RLS. The chapter included both theoretical and empirical analysis of the method, and showed that the new method is faster, more general, and in certain evaluation tasks, more reliable than the previously proposed method. Namely, in N-fold cross-validation, holding out basis vectors in each cross-validation repetition decreases the variance of the evaluation measure, and in tasks that do not fulfill the assumption of independently and identically distributed data, hold-out basis vectors are a necessity from the reliability point of view.

Chapter 5

Conclusions, Significance, and Future Work

5.1 Conclusions

In this doctoral dissertation, machine learning and clinical text were studied in order to support health information flow. The contributions related to clinical needs were a model of the ideal information flow, a model of the problems and challenges in reality, and a road map for the HLT development. Altogether five machine learning applications for clinical text were described. Their performance was evaluated in three practical cases. Also the associations between evaluation measures and methods were addressed. Finally, a new hold-out method for a particular learning algorithm was introduced. These contributions have received scientific recognition (e.g., a student encouragement award from Suominen et al. (2006), the best paper and presentation award from Suominen (2007), the third score in the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge (Computational Medicine Center, 2007)).

5.1.1 Clinical needs

For high-quality care, it is crucial that all the members of a multi-professional health care provider team and the patients themselves have the means to access, share, and utilize the gathered health records. Efficient information access is emphasized in the information-intensive and complex domain of health care, where clinicians are responsible for making decisions with substantial, or even life-and-death, impact on their patients' lives. In intensive care, these crucial decisions must be made extremely fast due to patients' critical status. The relevant data must be accessible in a timely and intelligible form, otherwise inabilities create risks to care safety and cost-effective health care administration.

A clear need for supporting the information flow of clinical narratives was indicated in this dissertation. When compared with the ideal (Figure 2.1), the flow from previous narratives is currently fragmented (Figure 2.5). This is explained by the overwhelming amount and topical-scattering of text, as well as unsystematic headings that complicate finding relevant information.

To address these fragments, the dissertation has introduced a road map for developing HLT (Figure 2.6) in order to create comprehensive solutions. The components of the road map support writing narratives, combining various sources of information, focusing attention on a particular topic, as well as searching and summarizing the content. These components take individual patient profiles into account and include proof-reading. Proof-reading contributes to text intelligibility to human-readers and performance of the other components.

5.1.2 Machine learning applications and their performance

The five machine learning applications described in this dissertation are related to the patient profile building, attention-focusing, and summarization components of the road map (Figure 2.6). The first two applications are binary classification and regression related to the practical case of topic labeling and relevance ranking. The next two applications are supervised and unsupervised multi-class classification for the practical case of topic segmentation and labeling. These four applications are tested with Finnish ICU nursing narratives. The fifth application is multi-label classification for the practical task of diagnosis coding. It is tested with English narratives written by physicians and radiologists. The performance of all these applications is promising.

The experiences gained in the method development give evidence of the following domain constraints:

1. a lack of annotated data and consequent need for minimally supervised methods,

2. a need for tailoring for highly specialized, telegraphic jargon,

3. a need for linguistic processing to reduce problems related to the sparse data,

4. the substantially small proportion of topically relevant text to irrelevant text that makes both the learning task and its performance evaluation challenging, and

5. the differing learning-task design and difficulty for the different topics and relevance-degrees

(Suominen and Salakoski, 2009). In addition, an interdisciplinary approach is needed for assuring the quality of the applications. In this dissertation, collaboration with nursing scientists, who are experts in clinical documentation and decision making, has enabled us to identify problems, whose solutions are of a substantial practical value. Information systems sciences connect these clinical needs and machine learning tasks, which, in turn, belong to the expertise of computer science.

5.1.3 Evaluation reliability

Emphasized demands on reliable solutions in health care has motivated the third research aspect of the dissertation: performance evaluation. Special care has been devoted to reviewing evaluation measures and justifying their use.

In order to address evaluation methods and their implementation, the dissertation has discussed hold-out techniques. In particular, taking dependences into account when holding out data and different aims for cross-validation have been addressed. Finally, a new hold-out method for sparse RLS has been introduced. This method is computationally efficient, contributes to evaluation reliability, and allows holding out multiple inputs simultaneously.

5.2 Computer science: significance and future work

Models are pervasive in science and are widely used in almost every area of business — from banking and insurance to health and medicine — to support decision making and improve human performance. Machine learning has made significant advances in abilities to construct realistic models, but the standard learning algorithms are not equipped to handle domain constraints. This dissertation has described the constraints related to health care HLT by addressing clinical needs; linguistic features; information search and diagnosis coding tasks; as well as reliable performance evaluation.

The dissertation has clarified performance evaluation and introduced a new evaluation method. Getting reliable performance evaluation results is difficult. Countless evaluation measures and methods exists for various machine learning tasks, but their relations are poorly understood. The value of such systematic study has been shown, for example, in Reid and Williamson (2009) in the case of binary classification. My dissertation has clarified the relations between evaluation measures. It has also introduced a new performance evaluation method that is particularly suitable for applications, where the data does not fulfill the assumption of independently and identically distributed data. This assumption is not often met with real-life

applications. When compared with previously proposed methods, the new one is faster and allows holding out multiple inputs simultaneously.

The topical information search approach discussed in this dissertation has the capability for scientific breakthroughs in automated step-wise summarization tailored for given purposes, especially in well-defined sub-domains such as Finnish ICU language. In machine learning, much of the ultimate goal of building systems able to communicate in human language remains out of reach for the current technology. For instance, fully automated summarization is an extremely challenging task, and the current solutions do not come close to meeting the quality requirements of clinical applications. In automated summarization, the first step is to specify the topics to be summarized. Human experts can do this, and automated content analysis can be used to support the task, for example, by identifying the most common topics. The dissertation addressed both of these view points. The second step is to find the relevant segments automatically. The dissertation developed these search tools by developing applications for automated topic segmentation, topic labeling, and relevance ranking. The topic segmentation, labeling, and ranking results can be used to ease writing summaries manually, or as an input for further automated summarization techniques. From the method development perspective, this thesis has contributed to the state-of-the-art in capabilities to segment short topics.

Future work should address the clinical need for minimally supervised methods. With LSA-HMM, more work is needed for the unsupervised selection of the model parameters. Furthermore, the use of the probabilistic interpretation of LSA might lead back to a proper HMM, which would open interesting directions for developing probabilistic models for sequence labeling. Another topic for further work would be to study the effect of various other methods providing unsupervised similarity measures (e.g., Probabilistic LSA (Hofmann, 1999) and Random Indexing (Kanerva et al., 2000)). Studying word similarities is likely to cast light on the content and structure of narratives which gives insights into the design of HLT applications too (Karlgren et al., 2008). When considering the topic labeling task alone, unsupervised techniques for topic-specific segment scoring may be fruitful in a multi-label classification sense (Buntine et al., 2005). Finally, a general way of modeling the topic *Other* is needed for applications where some text segments do not belong to any keyword-defined topic.

Finally, the connections between evaluation measures can be used in future work to understand how various machine learning algorithms theoretically relate to each other. By understanding these relations, task-specific constraints can be incorporated into learning. This study direction is emerging, but challenging; gaining significantly improved performance by incorporating constraints into training and model selection is challenging (Rakotomamonjy, 2004; Brefeld and Scheffer, 2005; Joachims, 2005). The previous

work includes, for example,

- changes in the evaluation measure (Rakotomamonjy, 2004; Brefeld and Scheffer, 2005; Joachims, 2005; Pahikkala et al., 2008a) or method (Bach et al., 2005; Pahikkala et al., 2008b),
- penalizing various errors differently (Elkan, 2001; Scott and Nowak, 2005; Han et al., 2008),
- capabilities for the user to specify the acceptable risk level beforehand (Geibel and Wysotzki, 2005; Defourny et al., 2008),
- exploiting prior knowledge about relationships between the data features and learning task (Mannor and Tsitsiklis, 2006; Kveton et al., 2008; Kotlowski and Slowinski, 2009), and
- as a remedy for data with missing values, semi-supervised machine learning techniques with exploitation of task-specific constraints (Chang et al., 2007).

This research will lead to better machine learning algorithms, prediction models, and decision-making tools.

5.3 Health care: significance and future work

The dissertation has clinical significance, which in the future, can contribute to individual patients, health care sciences, as well as the whole welfare of society. This research has produced new information about clinical documentation practices and HLT tailoring, as well as provided tools for supporting information flow that can next be piloted and developed further. The results of preliminary piloting confirm that health care professionals perceive the tools as useful (Lingsoft, 2008, 2009a). The pilot study has been conducted in the fall 2008, and it has included basic components for producing and using narratives (Figure 2.7): linguistic and stylistic proof-reading, domain-terminology building, and aid in understanding (i.e., linking to dictionaries and terminologies).

The dissertation has been conducted as a part of the larger *Louhi* (Text mining of patient records [Potilasasiakirjojen tekstin LOUHInta]) project[1] that belongs to the FinnWell — Future Healthcare Technology Programme of the Finnish Funding Agency for Technology and Innovation, Tekes (grants 40435/05, 40020/07). Through this project, the dissertation has a unique position of being able to establish a living link between health care service

[1] See `http://www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi/` [cited 2009 August 1].

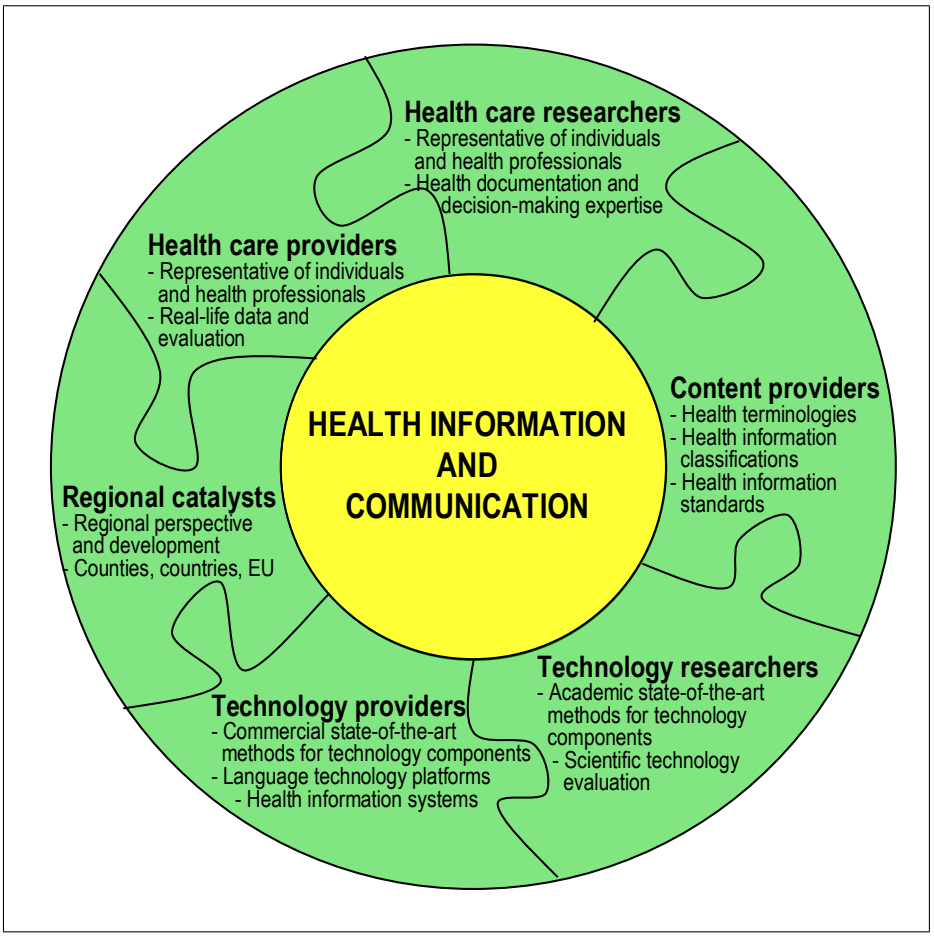


Figure 5.1: *Actors along the value chain of producing and using health information*

providers, health terminology developers, software houses in patient information systems and HLT, as well as research groups in health informatics.

The outcome of the collaboration has been a release of the first commercial proof-reading program for clinical Finnish (Lingsoft, 2009b) and an establishment of the consortium in 2008. This *IKITIK* (Information and language technology for health information and communication [Informaatio- ja KIeliteknologiaa TerveystIedon ja -Kommunikaation tueksi]) consortium[2] continues this dissertation work with an aim to support producing and using health information and communication by developing innovative, intelligent, state-of-the-art clinical information and language technology solutions. Its partners represent industry, academia, and health care service providers,

[2]See `http://www.ikitik.fi/` [cited 2009 August 1].

which promotes collaboration of all actors along the value chain of producing and using health information (Figure 5.1). The partners integrate the research results to commercial state-of-the-art platforms, and test the outcome in real environment providing end-user feedback.

IKITIK collaborates with the international *HEXAnord* (HEalth teXt Analysis network in the Nordic and Baltic countries; NordForsk project no. 9141) network[3]. The HEXAnord partners are the Danish Technical University, Denmark; University of Tartu, Estonia; University of Turku, Finland; Vytautas Magnus University, Lithuania; Norwegian University of Science and Technology, Norway; and KTH-Stockholm University, Sweden. In addition, international collaboration of IKITIK includes committed work in ScanBalt[4] activities in the Baltic Sea region as well as promoting European-level projects with governmental, academic, and industrial partners from Estonia, Finland, Hungary, the Netherlands, Sweden, and Switzerland.

More knowledge is needed about performance of the methods in clinical practice. In particular, evaluation of culture- and language-dependences deserves more attention in future work, because due to confidentiality of patient narratives, little is known about international applicability of existing solutions. A research plan and permits have already been established for me to carry out in an international collaboration comparative research on Finnish and Swedish patient records (Hospital District of Southwest Finland, research permit number 2/09 and 3/09; Ethical Committee of the Hospital District of Southwest Finland, record number 12.2.2009§66 and 12.2.2009§67; and Ethical Committee in Stockholm, EPN, record number 2008/5:2). This includes documenting style, structure, and content; environmental factors such as the patient information system used; and applying methods from one language to another. The Nordic health care system is known for its good quality, and hence its modeling is justified. This may enable transmitting the best Nordic health documentation practices and informatics internationally.

In the future, HLT can improve intelligibility of health information and in this way, support clinical decision making and empower patients cognitively. Patient documentation

- is a legal obligation,
- stands as an official proof of all necessary and sufficient information, which is needed for organizing, planning, performing and controlling good quality patient care,
- must be clear, intelligible, and correct, and

[3]See `http://dsv.su.se/en/research/ithealth/projects/hexanord` [cited 2009 November 20].

[4]See `http://www.scanbalt.org/` [cited 2009 November 20].

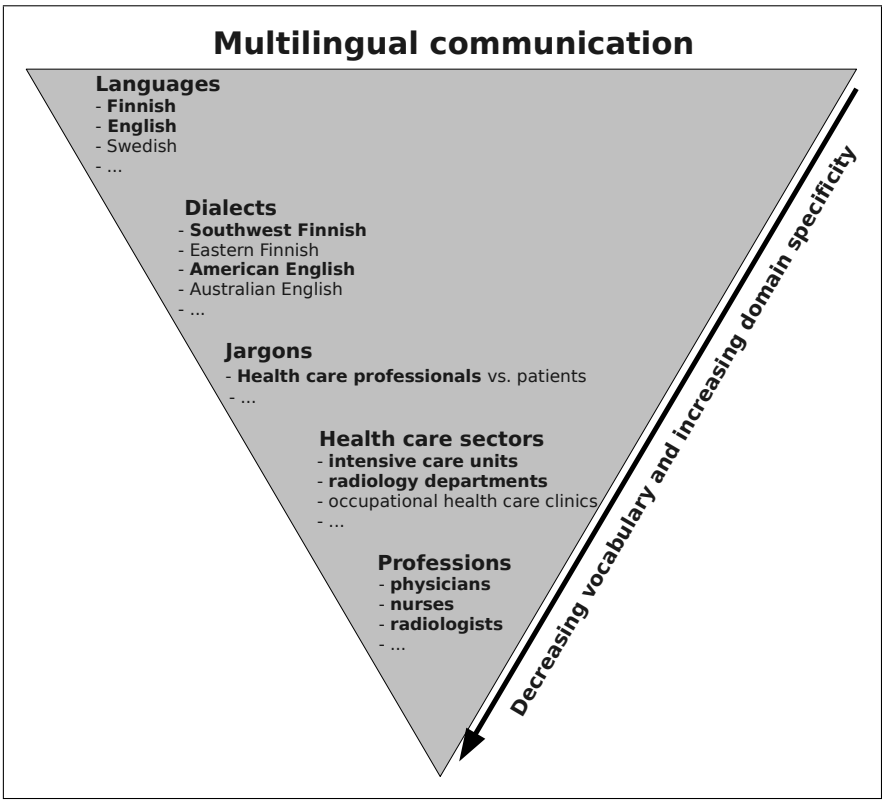


Figure 5.2: *Sub-languages and those discussed in this dissertation*

- should use only a generally known and widely accepted terminology and abbreviations

(Decree 99/2001 of the Ministry of Social and Health, Finland). This is an increasingly complex demand in the world of globalization and unit-specific practices (Section 2, Figures 2.4 and 5.2). Further, the knowledge stated narratives is not readily available for further automated analysis due to the ambiguous nature of human language, where the meaning of a sentence depends on its context and, on the other hand, a single meaning can be expressed in a number of equivalent ways. Currently, the records are produced and used by health care professionals. But individuals (i.e., patients or customers of health care services, or citizens) could use the technology themselves to ease accessing their health data, and in the future, even enter data to their records (Figure 1.2). In conclusion, methods resolving ambiguity as well as problems in grammar and understanding the content are needed to ease communication between health care professionals, improve

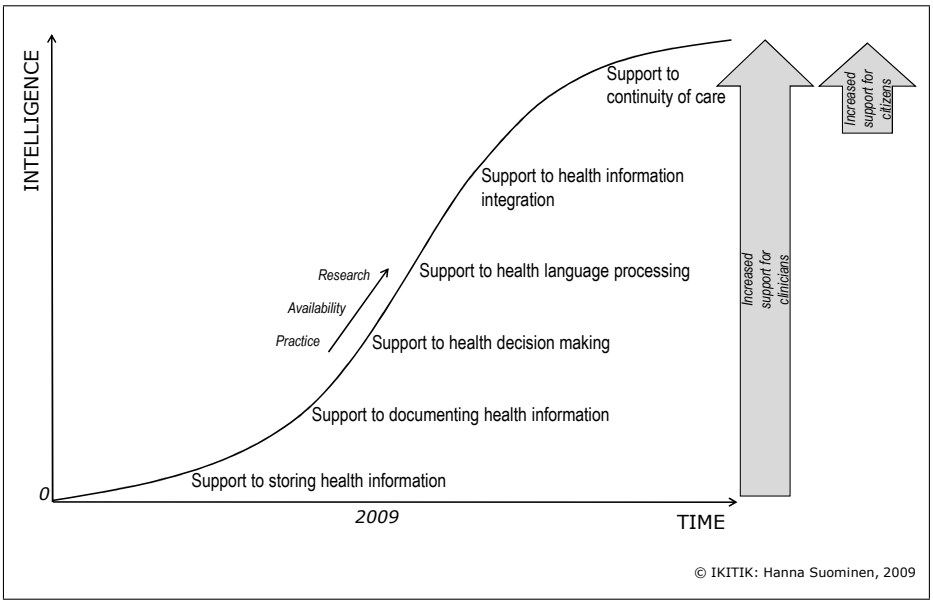


Figure 5.3: *Evolving electronic health records*

the performance of automated methods, as well as enhance patients' awareness regarding their health and encourages them in promoting health.

HLT tools have also the potential to enhance the quality and efficiency of care by improved access of health information. The tools will allow health care professionals more time for direct care. Further, supporting their decision making enables faster and more efficient identification and response to evolving health status trajectories. This has positive impacts on care outcomes, patient safety, as well as efficiency and profitability of health care services. Moreover, the HLT-assisted text summarization can take information needs of the target audience into account and extend the study from ICUs to other hospital wards. This intra-organizational information flow has been discussed, for example, in Ellingsen and Monteiro (2006) and Pinelle and Gutwin (2006).

The same technology will allow systematic analysis of masses of textual patient records, and use of this cumulative, knowledge-rich resource to enrich clinical evidence base. This enables better and more versatile usage of knowledge gained with previous patients, when caring for new patients, and consequently supports individualized care. In addition, these text mining capabilities can be utilized in science by creating practice-based evidence. Finally, ability to base care decisions on accurate, timely, and customized views on relevant information is likely to cause a paradigm shift in decision

making and administration in health care, analogously to what has already happened in modern private companies utilizing advanced business intelligence methods in their management and planning.

Health care service providers describe this electronic health record development using a model containing evolving generations (Figure 5.3). The first generation is the simplest, and it provides support to storing health information, that is, this write-only memory. The later ones are incorporated with increasing functionalities, including all previous capabilities. The second generation includes, in addition to storage, support to documenting health care information, and the third for decision making. This is the level of electronic health records that are currently used in health care, but HLT components still belong largely to research. These generations are based on Sensmeier (2003)and Handler (2004). Future generations, as increasing machine intelligence is gained, include first the HLT components (i.e., the fourth generation). The fifth generation electronic health records integrate data, in all formats, within all wards in one health care unit. The sixth generation supports continuity of care by combining all health data of individuals from all health care service providers, and it has the capabilities to capture past care experiences in order to use them in future care. In addition, individuals can supplement the data produced by health care professionals with their own notes and use their own electronic health records. The paradigm shift would be the result after six generations.

References

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., and Zhai, C. (2003). Challenges in information retrieval and language modeling: report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, september 2002. *SIGIR Forum*, 37(1):31–47.

Ambrosius, Huittinen, V.-M., Kari, A., Leino-Kilpi, H., Niinikoski, J., Ohtonen, M., Rauhala, V., Tammisto, T., and Takkunen, O. (1997). Suomen tehohoitoyhdistyksen eettiset ohjeet [The Ethical Guidelines of the Finnish Society of Intensive Care]. *Tehohoito [Journal of the Finnish Society of Intensive Care, FSIC]*, 15(2):165–72.

Anderson, B., Bross, I. D. J., and Sager, N. (1975). Grammatical compression in notes and records: Analysis and computation. *American Journal of Computational Linguistics*, 2(4):68–82.

Bach, F. R., Heckerman, D., and Horvitz, E. (2005). On the path to an ideal ROC curve: considering cost asymmetry in learning classifiers. In Cowell, R. and Ghahramani, Z., editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005*, pages 9–16. Society for Artificial Intelligence and Statistics.

Bakken, S., Hyun, S., Friedman, C., and Johnson, S. B. (2005). ISO reference terminology models for nursing: applicability for natural language processing of nursing narratives. *International Journal of Medical Informatics*, 74(1–3):615–622.

Banner, L. and Olney, C. M. (2009). Automated clinical documentation: does it allow nurses more time for patient care? *Journal of Critical Care*, 27(2):75–81.

Becker, H. (1972). Computerization of patho-histological findings in natural language. *Pathologia Europaea*, 7(2):193–200.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105.

Berman, J. J. (2002). Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine*, 26(1–2):25–36.

Blei, D. M. and Moreno, P. J. (2001). Topic segmentation with an aspect hidden Markov model. In Croft, W. B., Harper, D. J., Kraft, D. H., and Zobel, J., editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001*, pages 343–348. Association for Computing Machinery, New York, New York, USA.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl 1):D267–D270.

Bramsen, P., Deshpande, P., Lee, Y. K., and Barzilay, R. (2006). Finding temporal order in discharge summaries. *AMIA Annual Symposium Proceedings*, 2006:81–85.

Brefeld, U. and Scheffer, T. (2005). AUC maximizing support vector learning. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning*. Available from: `http://users.dsic.upv.es/~flip/ROCML2005/papers/brefeldCRC.pdf` [cited 2009 May 6].

Buntine, W., Löfström, J., Perttu, S., and Valtonen, K. (2005). Topic-specific link analysis using independent components for information retrieval. In Mladenic, D., Milic-Frayling, N., and Grobelnik, M., editors, *Association for the Advancement of Artificial Intelligence 2005 Workshop: Link Analysis*, AAAI Technical Report WS-05-07. AAAI Press, Menlo Park, California.

Castilla, A. C., Furuie, S. S., and Mendonça, E. A. (2007). Multilingual information retrieval in thoracic radiology: feasibility study. In Kuhn, K. A., Warren, J. R., and Leong, T.-Y., editors, *Proceedings of the 12th World Congress on Health (Medical) Informatics, MEDINFO 2007*, volume 129 of *Studies in Health Technology and Informatics*, pages 387–391. IOS Press, Amsterdam, the Netherlands.

Cawley, G. C. and Talbot, N. L. C. (2004). Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475.

Chang, M.-W., Ratinow, L., and Roth, D. (2007). Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL 2007*, pages 280–287. Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA.

Chang, T.-H. and Lee, C.-H. (2003). Topic segmentation for short texts. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation, PACLIC 17*, pages 159–165. COLIPS Publications, Singapore, Singapore.

Cheevakasemsook, A., Chapman, Y., Francis, K., and Davies, C. (2006). The study of nursing documentation complexities. *International Journal of Nursing Practice*, 12(6):366–374.

Chi, E. C., Sager, N., Tick, L. J., and Lyman, M. S. (1983). Relational data base modelling of free-text medical narrative. *Medical Informatics = Médecine et informatique*, 8(3):209–223.

Chinchor, N. and Sundheim, B. (1993). MUC-5 evaluation metrics. In *Proceedings of the Fifth Conference on Message Understanding, MUC 1993*, pages 69–78. Association for Computational Linguistics, Baltimore, Maryland, USA.

Cho, P. S., Taira, R. K., and Kangarloo, H. (2003). Automatic section segmentation of medical reports. *AMIA Annual Symposium Proceedings*, 2003:155–159.

Cios, K. J. and Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1–2):1–24.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(3):37–46.

Cohen, W. W. (1995). Fast effective rule induction. In Prieditis, A. and Russell, S., editors, *Proceedings of the Twelfth International Conference on Machine Learning, ICML 1995*, pages 115–123. Morgan Kaufmann, San Francisco, California, USA.

Collen, M. F. (1978). Patient data acquisition. *Medical Instrumentation*, 12(4):222–225.

Collier, N., Nazarenko, A., Baud, R., and Ruch, P. (2006). Recent advances in natural language processing for biomedical applications. *International Journal of Medical Informatics*, 75(6):413–417.

Computational Medicine Center (2007). *The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge*. Available from: `http://www.computationalmedicine.org/challenge` [cited 2009 May 6].

Cortes, C. and Mohri, M. (2004). AUC optimization vs. error rate minimization. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, Massachusetts, USA.

Crammer, K., Dredze, M., Ganchev, K., and Talukdar, P. P. (2007). Automatic code assignment to medical text. In Cohen, K. B., Demner-Fushman, D., Friedman, C., Hirschman, L., and Pestian, J., editors, *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136. Association for Computational Linguistics , Morristown, New Jersey, USA.

Currie, A. M., Fricke, T., Gawne, A., Johnston, R., Liu, J., and Stein, B. (2006). Automated extraction of free-text from pathology reports. *AMIA Annual Symposium Proceedings*, 2006:899.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Defourny, B., Ernst, D., and Wehenkel, L. (2008). Risk-aware decision making and dynamic programming. In *NIPS 2008 Workshop on Model Uncertainty and Risk in RL*. Available from: `http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2008/DEW08a` [cited 2009 May 6].

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Dorow, B. and Widdows, D. (2003). Discovering corpus-specific word senses. In Copestake, A. and Hajic, J., editors, *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2003*, pages 79–82. Association for Computational Linguistics, Morristown, New Jersey, USA.

Elkan, C. (2001). The foundations of cost-sensitive learning. In Nebel, B., editor, *Proceedings of the Seventeenth International Joint Conference on*

Artificial Intelligence, IJCAI 2001, pages 973–978. Morgan Kaufmann, San Francisco, California, USA.

Ellingsen, G. and Monteiro, E. (2006). Seamless integration: standardization across multiple local settings. *Computer Supported Cooperative Work*, 15(5–6):443–466.

Erdal, S. and Kamal, J. (2006). An indexing scheme for medical free text searches: a prototype. *AMIA Annual Symposium Proceedings*, 2006:918.

Fagerström, L., Rainio, A. K., Rauhala, A., and Nojonen, K. (2000). Validation of a new method for patient classification, the Oulu Patient Classification. *Journal of Advanced Nursing*, 31(2):481–490.

Farkas, R. and Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(Suppl 3):S10.

Fawcett, T. and Flach, P. (2005). A response to Webb and Ting's On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, 58(1):33–38.

Ferret, O. (2002). Using collocations for topic segmentation and link detection. In Tseng, S.-C., Chen, T.-E., and Liu, Y.-F., editors, *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002*, pages 1–7. Association for Computational Linguistics, New Brunswick, New Jersey, USA.

Friedlin, J. and McDonald, C. J. (2006). Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annual Symposium Proceedings*, 2006:925.

Friedman, C. and Hripcsak, G. (1999). Natural language processing and its future in medicine. *Academic Medicine*, 74(8):890–895.

Gabrieli, E. R. and Speth, D. J. (1986). Automated analysis of the discharge summary. *Journal of Clinical Computing*, 15(1):1–28.

Geibel, P. and Wysotzki, F. (2005). Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108.

Georgescul, M., Clark, A., and Armstrong, S. (2006). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *Proceedings of the Seventh SIGdial Workshop on Discourse and Dialogue*, pages 144–151. BPA Digital for the Association for Computational Linguistics, Burwood, Victoria, Australia.

Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2008). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. In Salakoski, T., Rebholz-Schuhmann, D., and Pyysalo, S., editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, SMBM 2008*, volume 51 of *TUCS General Publication*, pages 37–44. Turku Centre for Computer Science, Turku, Finland.

Glaser, S. R., Zamanou, S., and Hacker, K. (1987). Measuring and interpreting organizational culture. *Management Communication Quarterly*, 1(2):173–198.

Goldstein, I., Arzumtsyan, A., and Uzuner, Ö. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA Annual Symposium Proceedings*, 2007:279–283.

Goodwin, L., VanDyne, M., Lin, S., and Talbert, S. (2003). Data mining issues and opportunities for building nursing knowledge. *Journal of Biomedical Informatics*, 36(4–5):379–388.

Gruber, A., Rosen-Zvi, M., and Weiss, Y. (2007). Hidden topic Markov models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007*, volume 2 of *JMLR Workshop and Conference Proceedings*.

Gundlapalli, A. V., Olson, J., Smith, S. P., Baza, M., Hausam, R. R., Eutropius, L. J., Pestotnik, S. L., Duncan, K., Staggers, N., Pincetl, P., and Samore, M. H. (2007). Hospital electronic medical record-based public health surveillance system deployed during the 2002 Winter Olympic Games. *American Journal of Infection Control*, 35(3):163–171.

Hakes, B. and Whittington, J. (2008). Assessing the impact of an electronic medical record on nurse documentation time. *Journal of Critical Care*, 26(4):234–241.

Han, M., Chen, D., and Sun, Z. (2008). Analysis to Neyman-Pearson classification with convex loss function. *Analysis in Theory and Applications*, 24(1):18–28.

Handler, T. J. (2004). *The 2004 Gartner Computer-based Patient Record System Generation Model*. Gartner, Stamford, Connecticut, USA. Strategic analysis report R-21-6592.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hanson, C. W. and Marshall, B. E. (2001). Artificial intelligence applications in the intensive care unit. *Critical Care Medicine*, 29(2):427–435.

Hearst, M. A. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heldt, T., Long, B., Verghese, C., Szolovits, P., and Mark, R. G. (2006). Integrating data, models, and reasoning in critical care. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006*, pages 350–353. IEEE Service Center, New York City, New York, USA.

Hellesø, R. (2006). Information handling in the nursing discharge note. *Journal of Clinical Nursing*, 15(1):11–21.

Hirschman, L., Grishman, R., and Sager, N. (1976). From text to structured information: automatic processing of medical reports. In *American Federation of Information Processing Societies: 1976 National Computer Conference*, volume 45 of *AFIPS Conference Proceedings*, pages 267–275.

Hirschman, L. and Mani, I. (2003). Evaluation. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 414–429. Oxford University Press, Oxford, New York, USA.

Hirschman, L. and Thompson, H. S. (1997). Overview of evaluation in speech and natural language processing. In Cole, R., editor, *Survey of the State of the Art in Human Language Technology*, pages 409–414. Cambridge University Press, New York, New York, USA.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence, UAI 1999*, pages 289–296. Morgan Kaufmann, San Francisco, California, USA.

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., and Melton, G. B. (2007). Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia. *Computers in Biology and Medicine*, 37(3):296–304.

Huang, Y., Lowe, H. J., Klein, D., and Cucina, R. J. (2005). Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS Specialist Lexicon. *Journal of the American Medical Informatics Association: JAMIA*, 12(3):275–285.

Hyun, S. and Bakken, S. (2006). Toward the creation of an ontology for nursing document sections: mapping section headings to the LOINC semantic model. *AMIA Annual Symposium Proceedings*, 2006:364–368.

Jancsary, J. and Matiasek, J. (2008). Revealing the structure of medical dictations with conditional random fields. In Lapata, M. and Ng, H. T., editors, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008*, pages 1–10. Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine learning, ICML 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 377–384. Association for Computing Machinery, New York, New York, USA.

Kanerva, P., Kristofersson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society, CogSci 2000*, page 1036. University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Kärkkäinen, O. and Eriksson, K. (2003). Evaluation of patient records as part of developing a nursing care classification. *Journal of Clinical Nursing*, 12(2):198–205.

Karlgren, J., Holst, A., and Sahlgren, M. (2008). Filaments of meaning in word space. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R. W., editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 531–538. Springer, Berlin / Heidelberg, Germany.

Kaustinen, T. (1995). *Hoitoisuusluokituksen kehittäminen ja arviointi [Development and Evaluation of Patient Classification]*. PhD thesis, University of Oulu, Department of Nursing Science, Oulu, Finland.

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

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Mellish, C., editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 1995*, volume 2, pages 1137–1143. Morgan Kaufmann, San Mateo, California, USA.

Koskenniemi, K. (1983). Two-level model for morphological analysis. In Bundy, A., editor, *Proceedings of the Eighth International Joint Con-*

ference on Artificial Intelligence, IJCAI 1983, pages 683–685. Morgan Kaufmann, San Francisco, California, USA.

Kotlowski, W. and Slowinski, R. (2009). Rule learning with monotonicity constraints. In Danyluk, A. P., Bottou, L., and Littman, M. L., editors, *Proceeding of the Twenty-Sixth International Conference on Machine Learning, ICML 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 537–544. Association for Computing Machinery, New York, New York, USA.

Kveton, B., Yu, J. Y., Theocharaus, G., and Mannor, S. (2008). Online learning with expert advice and finite-horizon constraints. In Fox, D. and Gomes, C. P., editors, *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, pages 331–336. AAAI Press, Menlo Park, California, USA.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Brodley, C. E. and Danyluk, A. P., editors, *Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001*, pages 282–289. Morgan Kaufmann, San Francisco, California, USA.

Laippala, V., Ginter, F., Pyysalo, S., and Salakoski, T. (2009). Towards automated processing of clinical Finnish: sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics*, 78(12):e7–e12.

Lauri, S. and Salanterä, S. (2002). Developing an instrument to measure and describe clinical decision making in different nursing fields. *Journal of Professional Nursing*, 18(2):93–100.

Lingsoft (2008). *Monikielityöpöydän TYKS-koekäytön testaustulokset [The evaluation results of the multilingual human language technology platform in the Turku University Hospital].* Internal project report.

Lingsoft (2009a). *IKITIK kehittää parempaa hoitoa ja elämänhallintaa ymmärrettävällä ja laadukkaalla kielellä [The IKITIK consortium develops better care and quality of life via intelligible and good language].* Press release, 2009 March 26, `http://www.lingsoft.fi/?doc_id=433&lang=fi` [cited 2009 August 1].

Lingsoft (2009b). *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` [cited 2009 August 1].

Lovis, C., Baud, R. H., and Planche, P. (2000). Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics*, 58–59(1):101–110.

Lundgrén-Laine, H. and Suominen, H. (2009). Mistä näitä tietoja oikein tulee? kieliteknologialla tehohoidon tiedot hallintaan [Where do all these data come from? Solving problems in knowledge management of intensive care data via human language technology]. *Tehohoito [Journal of the Finnish Society of Intensive Care, FSIC]*, 27(2):102–104.

Lundgrén-Laine, H., Suominen, H., Kontio, E., Salanterä, S., and Salakoski, T. (2009). Intensive care admission and discharge — critical decision-making points. In Saranto, K., Flatley Brennan, P., Park, H.-A., Tallberg, M., and Ensio, A., editors, *Connecting Health and Humans. Proceedings of NI2009, the 10th International Congress of Nursing Informatics*, volume 146 of *Studies in Health Technology and Informatics*, pages 358–361. IOS Press, Amsterdam, the Netherlands.

Lussier, Y. A., Shaginai, L., and Friedman, C. (2000). Automating ICD-9-CM encoding using medical language processing: a feasibility study. *AMIA Annual Symposium Proceedings*, 2000:1072.

Mannor, S. and Tsitsiklis, J. N. (2006). Online learning with constraints. In Simon, H. U. and Lugosi, G., editors, *Proceedings of the 19th Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 529–543. Springer, Berlin / Heidelberg, Germany.

Manor-Shulman, O., Beyene, J., Frndova, H., and Parshuram, C. (2008). Quantifying the volume of documented clinical information in critical illness. *Journal of Critical Care*, 23(2):245–250.

Mendonça, E. A., Haas, J., Shagina, L., Larson, E., and Friedman, C. (2005). Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*, 38(4):314–321.

Meystre, S. and Haug, P. J. (2006a). Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. *AMIA Annual Symposium Proceedings*, 2002:554–558.

Meystre, S. and Haug, P. J. (2006b). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic

health record: a review of recent research. In Kulikowski, C. A. and Geissbühler, A., editors, *IMIA Yearbook of Medical Informatics 2008*, pages 128–144. Schattauer GmbH, Stuttgart, Germany.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE Publications, Thousand Oaks, California, USA, 2nd edition.

Moisio, M. A. (2000). *A Guide to Health Care Insurance Billing*. Thomson Delmar Learning, Clifton Park, New York, USA.

Mullin, M. and Sukthankar, R. (2000). Complete cross-validation for nearest neighbor classifiers. In Langley, P., editor, *Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000*, pages 639–646. Morgan Kaufmann Publishers, San Francisco, California, USA.

National Advisory Board on Research Ethics (2002). *Good Scientific Practice and Procedures for Handling Misconduct and Fraud in Science*. Available from: `http://www.tenk.fi/ENG/HTK/index.htm` [cited 2009 February 19].

National Center for Health Statistics (2007). *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*. Available from: `http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm` [cited 2009 May 6].

Norman, G. R. and Streiner, D. L. (2000). *Biostatistics: the Bare Essentials*. B. C. Decker Inc., Hamilton, Ontario, Canada, 2nd edition.

Ogren, P. V. (2006). Knowtator: a Protégé plug-in for annotated corpus construction. In Moore, R. C., Bilmes, J. A., Chu-Carroll, J., and Sanderson, M., editors, *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 273–275. Association for Computational Linguistics, New York, New York, USA.

Pahikkala, T. (2008). *New Kernel Functions and Learning Methods for Text and Data Mining*. PhD thesis, TUCS Dissertation 103, Turku Centre for Computer Science, Turku, Finland.

Pahikkala, T., Airola, A., Suominen, H., Boberg, J., and Salakoski, T. (2008a). Efficient AUC maximization with regularized least-squares. In Holst, A., Kreuger, P., and Funk, P., editors, *Proceedings of the Tenth Scandinavian Conference on Artificial Intelligence, Frontiers in Artificial Intelligence and Applications, SCAI 2008*, volume 119, pages 12–19. IOS Press, Amsterdam, the Netherlands.

Pahikkala, T., Boberg, J., and Salakoski, T. (2006). Fast n-fold cross-validation for regularized least-squares. In Honkela, T., Raiko, T., te la, J. K., and Valpola, H., editors, *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence, SCAI 2006*, pages 83–90. Otamedia, Espoo, Finland.

Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2008b). Transductive ranking via pairwise regularized least-squares. In Frasconi, P., Kersting, K., and Tsuda, K., editors, *Proceedings of the 5th International Workshop on Mining and Learning with Graphs, MLG 2007*, pages 175–178. Florence, Italy, 2007 August 1–3.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Järvinen, J., and Boberg, J. (2009). An efficient algorithm for learning to rank from preference graphs. *Machine Learning*, 75(1):129–165.

Pakhomov, S., Weston, S. A., Jacobsen, S. J., Chute, C. G., Meverden, R., and Roger, V. L. (2007a). Electronic medical records for clinical research: application to the identification of heart failure. *American Journal of Managed Care*, 13(6 Part 1):281–288.

Pakhomov, S. S., Hemingway, H., Weston, S. A., Jacobsen, S. J., Rodeheffer, R., and Roger, V. L. (2007b). Epidemiology of angina pectoris: role of natural language processing of the medical record. *American Heart Journal*, 153(4):666–673.

Pakhomov, S. V., Buntrock, J. D., and Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association: JAMIA*, 13(5):516–525.

Pentz, J. F., Wilcox, A. B., and Hurdle, J. F. (2007). Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*, 40(2):174–182.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., and Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In Cohen, K. B., Demner-Fushman, D., Friedman, C., Hirschman, L., and Pestian, J., editors, *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, Morristown, New Jersey, USA.

Pevzner, L. and Hearst, M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Pierpont, G. L. and Thilgen, D. (1995). Effect of computerized charting on nursing activity in intensive care. *Critical Care Medicine*, 23(6):1067–1073.

Pinelle, D. and Gutwin, C. (2006). Loose coupling and healthcare organizations: deployment strategies for groupware. *Computer Supported Cooperative Work*, 15(5–6):537–572.

Poggio, T. and Smale, S. (2003). The mathematics of learning: dealing with data. *Notices of the American Mathematical Society (AMS)*, 50(5):537–544.

Ponte, J. M. and Croft, W. B. (1997). Text segmentation by topic. In Peters, C. and Thanos, C., editors, *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, ECDL 1997*, volume 1324 of *Lecture Notes in Computer Science*, pages 113–125. Springer, Berlin / Heidelberg, Germany.

Porter, M. and Boulton, R. (2006). *The Snowball Stemming Algorithm*. Available from: `http://snowball.tartarus.org/algorithms/finnish/stemmer.html` [cited 2006 January 1].

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rakotomamonjy, A. (2004). Optimizing area under ROC curve with SVMs. In Hernández-Orallo, J., Ferri, C., Lachiche, N., and Flach, P. A., editors, *Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence, ROCAI 2004*, pages 71–80. Valencia, Spain, 2004 August 22.

Reid, M. and Williamson, R. (2009). *Information, Divergence and Risk for Binary Experiments*. Available from: `http://arxiv.org/abs/0901.0356` [cited 2009 May 1].

Rifkin, R. (2002). *Everything Old Is New Again: a Fresh Look at Historical Approaches in Machine Learning*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Ruch, P., Baud, R., and Geissbühler, A. (2003). Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial Intelligent in Medicine*, 29(1–2):169–184.

Saeh, J. C., Lyne, P. D., Takasaki, B. K., and Cosgrove, D. A. (2005). Lead hopping using SVM and 3D pharmacophore fingerprints. *Journal of Chemical Information and Modeling*, 45(4):1122–1133.

Sætre, R., Sagae, K., and Tsujii, J. (2007). Syntactic features for protein-protein interaction extraction. In Baker, C. J. O. and Su, J., editors, *Short Paper Proceedings of the 2nd International Symposium on Languages in Biology and Medicine, LBM 2007*, volume 319 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sarawagi, S. and Cohen, W. W. (2005). Semi-Markov conditional random fields for information extraction. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 1185–1192. MIT Press, Cambridge, MA.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, R., editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 416–426. Springer, Berlin, Germany.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Scott, C. and Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819.

Sensmeier, J. (2003). Advancing the state of data integration in healthcare. *Journal of Healthcare Information Management*, 17(4):58–61.

Sessler, C. N., Grap, M. J., and Ramsay, M. A. E. (2008). Evaluating and monitoring analgesia and sedation in the intensive care unit. *Critical Care*, 12(Suppl 3):S2.

Shah, A. D. and Martinez, C. (2006). An algorithm to derive a numerical daily dose from unstructured text dosage instructions. *Pharmacoepidemiology and Drug Safety*, 15(3):161–166.

Shapiro, A. R. (1983). Exploratory analysis of the medical record. *Medical Informatics = Médecine et informatique*, 8(3):163–171.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, New York, USA, 2nd edition.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18.

Smith, K., Smith, V., Krugman, M., and Oman, K. (2005). Evaluating the impact of computerized clinical documentation. *CIN: Computers, Informatics, Nursing*, 23(3):132–138.

Smola, A. J. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In Langley, P., editor, *Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000*, pages 911–918. Morgan Kaufmann Publishers, San Francisco, California, USA.

Snyder-Halpern, R., Corcoran-Perry, S., and Narayan, S. (2001). Developing clinical practice environments supporting the knowledge work of nurses. *Computers in Nursing*, 19(1):17–23, quiz 24–26.

Society of Critical Care Medicine (2007). *What is Critical Care?* Available from: `http://www.myicucare.org/Critical_Care_Questions/Pages/default.aspx` [cited 2009 April 3].

Spärck Jones, K. and Galliers, J. (1996). Evaluating natural language processing systems: an analysis and review. Volume 1083 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Germany.

Stakes (2008). *The Statistical Yearbook on Social Welfare and Health Care 2008*. Yliopistopaino, Helsinki, Finland.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17).

Suominen, H. (2007). Kieliteknologian menetelmien soveltaminenpotilasdokumentaation hyödyntämiseen [Applying human language technology for patient documents]. In Koskinen, M. and Jauhiainen, E., editors, *Tietojenkäsittelytieteen päivät 2007 [The Computer Science Convention 2007]*, volume TU-25 of *Tutkimuksia, Tietojenkäsittelytieteiden julkaisuja [Studies, Computer science publications]*, pages 46–50. Jyväskylä University Press, Jyväskylä, Finland.

Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2006). Theoretical considerations of ethics in text mining of nursing documents. In Park, H. A., Murray, P., and Delaney, C., editors, *Consumer-Centered Computer-Supported Care for Healthy People, Proceedings of NI2006, The 9th International Congress on Nursing Informatics*, volume 122 of *Studies in Health Technology and Informatics*, pages 353–364. IOS Press, Amsterdam, the Netherlands.

Suominen, H., Lundgrén-Laine, H., Salanterä, S., Karsten, H., and Salakoski, T. (2009a). Information flow in intensive care narratives. In Chen, J., Chen, C., Ely, J., Hakkani-Tr, D., He, J., Hsu, H.-H., Liao, L.,

Liu, C., Pop, M., Ranganathan, S., Reddy, C. K., Ruan, J., Song, Y., Tseng, V. S., Ungar, L., Wu, D., Wu, Z., Xu, K., Yu, H., and Zelikovsky, A., editors, *Proceedings IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBM 2009*, pages 325–330. Institute of Electrical and Electronics Engineers, Los Alamitos, California, USA.

Suominen, H., Lundgrén-Laine, H., Salanterä, S., and Salakoski, T. (2009b). Evaluating pain in intensive care. In Saranto, K., Flatley Brennan, P., Park, H.-A., Tallberg, M., and Ensio, A., editors, *Connecting Health and Humans. Proceedings of NI2009, the 10th International Congress of Nursing Informatics*, volume 146 of *Studies in Health Technology and Informatics*, pages 191–196. IOS Press, Amsterdam, the Netherlands.

Suominen, H., Pyysalo, S., Ginter, F., and Salakoski, T. (2008a). Automated text segmentation and topic labeling of clinical narratives. In Karsten, H., Back, B., Salakoski, T., Salanterä, S., and Suominen, H., editors, *Proceedings of the First Conference on Text and Data Mining of Clinical Documents, Louhi 2008*, volume 52 of *TUCS General Publication*, pages 99–103. Turku Centre for Computer Science, Turku, Finland.

Suominen, H., Pyysalo, S., Hiissa, M., Ginter, F., Liu, S., Marghescu, D., Pahikkala, T., Back, B., Karsten, H., and Salakoski, T. (2008b). Performance evaluation measures for text mining. In Song, M. and Wu, Y.-F., editors, *Handbook of Research on Text and Web Mining Technologies*, volume 2, pages 724–747. IGI Global, Hershey, Pennsylvania, USA.

Suominen, H. and Salakoski, T. (2009). Information retrieval and personal health records: user needs and domain tailoring. In *SPIRE 2009 Workshop on Task-based Information Access*. The Swedish Institute of Computer Science, Kista, Sweden. In Press.

Suominen, H. J., Lehtikunnas, T., Hiissa, M., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2005). Natural language processing for nursing documentation. In Fonseca, J., editor, *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare, CIMED 2005*, pages 147–54, Costa da Caparica, Portugal, 2005 June 29 – July 01.

Suominen, H. J., Lundgrén-Laine, H. K., Perttilä, J. T., Salakoski, T. I., and Salanterä, S. M. (2008c). Tehohoidon elektroniset potilasasiakirjat — hyödyntämätön voimavara [Electronic intensive care patient records — an untapped resource]. In *Valtakunnalliset Lääkäripäivät 2008 [The Finnish Medical Convention 2008]*, Helsinki, Finland, 2008 January 6–10.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300.

Tange, H. (1996). How to approach the structuring of the medical record? Towards a model for flexible access to free text medical data. *International Journal of Biomedical Computing*, 42(1–2):27–34.

Tange, H., Hasman, A., de Vries Robbe, P., and Schouten, H. (1997). Medical narratives in electronic medical records. *International Journal of Medical Informatics*, 46(1):7–29.

Turchin, A., Kolatkar, N. S., Grant, R. W., Makhni, E. C., Pendergrass, M. L., and Einbinder, J. S. (2006). Using regular expression to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association: JAMIA*, 13(6):691–695.

US Department of Health & Human Services (1996). *Health Insurance Portability and Accountability Act of 1996, HIPAA 1996.* Available from: `http://www.cms.hhs.gov/HIPAAGenInfo/Downloads/HIPAALaw.pdf` [cited 2009 November 18].

Velupillai, S., Dalianis, H., Hassel, M., and Nilsson, G. (2009). Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):e19–e26.

Walsh, S. H. (2004). The clinician's perspective on electronic health records and how they can affect patient care. *British Medical Journal (Clinical Research Edition)*, 328(7449):1184–1187.

Ward, N. S., Snyder, J. E., Ross, S., Haze, D., and Levy, M. M. (2004). Comparison of a commercially available clinical information system with other methods of measuring critical care outcomes data. *Journal of Critical Care*, 19(1):10–15.

Weaver, C. A., Warren, J. J., and Delaney, C. (2005). Bedside, classroom and bench: collaborative strategies to generate evidence-based knowledge for nursing practice. *International Journal of Medical Informatics*, 74(11–12):989–999.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Yamron, J. P., Carp, I., Gillick, L., Lowe, S., and van Mulbregt, P. (1998). A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998*, volume 1, pages 333–336. Institute of Electrical and Electronic Engineers Signal Processing Society, Piscataway, New Jersey, USA.

Young, D. A. (1982). Language and the brain: implications from new computer models. *Medical Hypotheses*, 9(1):55–70.

Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., and Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 26(6):30.

Zhou, L., Tao, Y., Cimino, J. J., Chen, E. S., Liu, H., Lussier, Y. A., Hripcsak, G., and Friedman, C. (2006). Terminology model discovery using natural language processing and visualization techniques. *Journal of Biomedical Informatics*, 39(6):626–636.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375.

Part II

Publication Reprints

Paper I

Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2007). Applying language technology to nursing documents: pros and cons with a focus on ethics. *International Journal of Medical Informatics*, 76(S2):S293–S301.

Extends:
Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., and Salanterä, S. (2006). Theoretical considerations of ethics in text mining of nursing documents. In Park, H.-A., Murray, P., and Delaney, C., editors. *Consumer-Centered Computer-Supported Care for Healthy People. Proceedings of NI2006, the 9th International Congress of Nursing Informatics*, volume 122 of *Studies in Health Technology and Informatics*, pages 353–364. IOS Press, Amsterdam, the Netherlands. (Student encouragement award)

Paper II

Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2007). Towards automated classification of intensive care nursing narratives. *International Journal of Medical Informatics*, 76(S3):S362–S368.

Extends:
Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2006). Towards automated classification of intensive care nursing narratives. In Hasman, A., Haux, R., Van Der Lei, J., De Clercq E., and Roger France, F. H., editors. *Ubiquity: Technologies for Better Health in Aging Societies. Proceedings of MIE 2006, the 20th International Congress of the European Federation of Medical Informatics*, volume 124 of *Studies in Health Technology and Informatics*, pages 789–794. IOS Press, Amsterdam, the Netherlands.

Paper III

Suominen, H., Pahikkala, T., Hiissa, M., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., and Salakoski, T. (2006). Relevance ranking of intensive care nursing narratives. In Gabrys, B., Howlett, R. J, and Jain, L. C., editors. *Knowledge-Based Intelligent Information and Engineering Systems. Proceedings of the 10th International Conference, KES 2006, Part I*, volume 4251 of *Lecture Notes in Computer Science*, pages 720–727. Springer, Berlin / Heidelberg, Germany.

Paper IV

Ginter, F., Suominen, H., Pyysalo, S., Salakoski, T. (2009). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application. *International Journal of Medical Informatics*, 78(12):e1–e6.

Extends:
Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2008). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application. In Salakoski, T., Rebholz-Schuhmann, D., and Pyysalo, S., editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, SMBM 2008*, number 51 of *TUCS General Publication*, pages 37–44. Turku Centre for Computer Science, Turku, Finland.
and
Suominen, H., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). Automated text segmentation and topic labeling of clinical narratives. In Karsten, H., Back, B., Salakoski, T., Salanterä, S., and Suominen, H., editors. *Proceedings of the First Conference on Text and Data Mining of Clinical Documents, Louhi 2008*, number 52 of *TUCS General Publication*, pages 99–103. Turku Centre for Computer Science, Turku, Finland.

Paper V

Suominen, H., Ginter, F., Pyysalo, S., Airola, A., Pahikkala, T., Salanterä, S., and Salakoski, T. (2008). Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In Hauskrecht, M., Schuurmans, D., and Szepesvari, C., editors. *Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications*, 2008 July 9, Helsinki, Finland.

Paper VI

Suominen, H., Pahikkala, T., and Salakoski, T. (2008). Critical points in assessing learning performance via cross-validation. In Honkela, T., Pöllä, M., Paukkeri, M.-S., Simula, O., editors. *Proceedings of the 2nd International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR 2008*, pages 9–22. Multiprint, Espoo, Finland.

Paper VII

Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2009). Efficient cross-validation algorithms for sparse regularized least-squares. Unpublished manuscript.

Extends:
Pahikkala, T., Suominen, H., Boberg, J., and Salakoski, T. (2009). Efficient hold-out for subset of regressors. In Kolehmainen, M., editor. *Adaptive and Natural Computing Algorithms, 9th International Conference, ICANNGA 2009, Revised Selected Papers*, volume 5495 of *Lecture Notes in Computer Science*, pages 350–359. Springer, Berlin / Heidelberg, Germany.

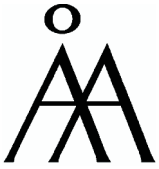
Turku Centre for Computer Science
TUCS Dissertations

91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics



Åbo Akademi University
- Department of Information Technologies

Turku School of Economics
- Institute of Information Systems Sciences

ISBN 978-952-12-2375-4
ISSN 1239-1883

Hanna Suominen

Machine Learning and Clinical Text: Supporting Health Information Flow

Hanna Suominen

Machine Learning and Clinical Text: Supporting Health Information Flow