TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

APPLICATIONS OF NEURAL LANGUAGE REPRESENTATIONS IN BIOMEDICAL AND CLINICAL TEXT CLASSIFICATION AND NAMED ENTITY RECOGNITION

Kai Hakala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 39 | TECHNICA – INFORMATICA | TURKU 2024

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

APPLICATIONS OF NEURAL LANGUAGE REPRESENTATIONS IN BIOMEDICAL AND CLINICAL TEXT CLASSIFICATION AND NAMED ENTITY RECOGNITION

Kai Hakala

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 39 | TECHNICA – INFORMATICA | TURKU 2024

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral Programme in Technology

Supervised by

Professor Filip Ginter
Department of Computing
Faculty of Technology
University of Turku, Finland

Professor Tapio Salakoski
Department of Mathematics and Statistics
Faculty of Science
University of Turku, Finland

Reviewed by

Associate Professor, Øystein Nytrø
UiT The Arctic University of Norway, Norway

Dr. Martin Duneld
Stockholm University, Sweden

Opponent

Professor, Hanna Suominen
Australian National University, Australia

ISBN 978-951-29-9725-1 (PRINT)
ISBN 978-951-29-9726-8 (PDF)
ISSN 2736-9390 (PRINT)
ISSN 2736-9684 (ONLINE)
Painosalama, Turku, Finland, 2024

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
HAKALA, KAI: Applications of neural language representations in biomedical and clinical text classification and named entity recognition
Doctoral dissertation, 139 pp.
Doctoral Programme in Technology
June 2024

ABSTRACT

The abundance of biomedical literature and clinical care documentation creates enormous challenges for the research community as well as for clinical decision-making due to the information overload and quality concerns. Additionally the resources available for providing clinical care are diminished by to the documentation burden. This naturally leads to the need for efficient tools in producing, standardizing, structuring and abstracting the textual information available.

Over the past decade, language technology has taken drastic leaps in modelling the structural and semantic aspects of language, in particular in the form of neural language models. These models rely on transfer learning in the form of being initially pretrained on large quantities of unannotated text, substantially reducing the amount of needed task-specific training data in the subsequent finetuning phase.

This thesis explores the use of pretrained neural networks for clinical and biomedical text mining, mostly in the form of text classification and named entity recognition. The emphasis is on thorough evaluation of neural network based models in selected text mining tasks, spanning across clinical care documentation, biomedical literature and social media as well as three different languages: namely English, Spanish and most importantly Finnish. The wide range of tasks provides a good overview on the applicability of neural transfer learning in the clinical domain.

The results suggest that the developed methods are able to reach performance levels comparable to domain experts, warranting the use of neural methods in real-world applications. Moreover, this work demonstrates the efficiency of multilingual and cross-domain transfer learning on clinical text mining, with cross-domain methods surpassing the performance of the domain-specific baselines. The method development and evaluation work is extended with preliminary analysis of the internal representations extracted from the neural models. This study illustrates a secondary use-case of neural language representations in the data-driven refinement of medical ontologies.

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietojenkäsittelytiede
HAKALA, KAI: Applications of neural language representations in biomedical and clinical text classification and named entity recognition
Väitöskirja, 139 s.
Teknologian tohtoriohjelma
kesäkuu 2024

TIIVISTELMÄ

Biolääketieteen tutkimuskirjallisuuden ja hoitotyöstä muodostuvan dokumentaation runsauden aiheuttama informaatiokuormitus ja dokumentaation laadulliset ongelmat tuottavat haasteita biolääketieteen tutkimusyhteisölle sekä vaikeuttavat kliinistä päätöksentekoa. Lisäksi dokumentoinnin luoma taakka vaikeuttaa resurssien kohdentamista varsinaiseen hoitotyöhön. Näistä syistä lääketieteellisen tekstin tuottamiseen, standardointiin, rakenteellistamiseen ja tiivistämiseen kehitettyjen työkalujen tarve on korostunut.

Viime vuosikymmenen aikana kieliteknologia on ottanut huomattavia harppauksia kielen rakenteen ja merkitysten mallintamisessa. Suuressa roolissa tässä kehityksessä ovat olleet neuroverkkoihin perustuvat kielimallit. Nämä mallit tukeutuvat siirto-oppimiseen, jossa ne esikoulutetaan suurilla määrillä raakatekstiä. Tällainen esikoulutus vähentää huomattavasti tehtäväkohtaisen koulutusdatan määrää hienosäätövaiheessa.

Väitöstutkimukseni tarkastelee esikoulutettujen neuroverkkojen käyttöä kliinisessä ja biolääketieteellisessä tekstinlouhinnassa etenkin tekstinluokittelun ja nimettyjen entiteettien tunnistamisen muodossa. Painopiste on neuroverkkoihin perustuvien mallien perusteellisessa arvioinnissa valituissa tekstinlouhintatehtävissä, jotka kattavat kliinisen hoidon dokumentaation, biolääketieteellisen kirjallisuuden sekä sosiaalisen median. Tarkasteltavat tekstilähteet muodostuvat englannin-, espanjan- sekä ennen kaikkea suomenkielisistä sisällöistä. Tehtävien laaja kirjo antaa hyvän yleiskuvan siirto-oppimisen sovellettavuudesta kliinisen tekstin koneelliseen tulkintaan.

Tutkimukseni tulokset osoittavat, että kehitettyjen menetelmien tarkkuus on tarkastelluissa tehtävissä verrattavissa hoitotyön asiantuntijoihin, mikä tukee vastaavien menetelmien käyttöönottoa aidoissa työympäristöissä. Lisäksi tutkimukseni osoittaa monikielisen ja toimialariippumattoman siirto-oppimisen tehokkuuden kliinisessä tekstinlouhinnassa, sillä tutkimuksessa kehitetyt toimialariippumattomat menetelmät ylittävät toimialakohtaisten verrokkimenetelmien suorituskyvyn. Menetelmien kehittämisen ja arvioinnin lisäksi tutkimuksessa tarkastellaan neuroverkkomallien sisäisiä kielen esityksiä. Tämä osuus havainnollistaa neuroverkkomallien kieliesitysten toisiokäyttöä lääketieteellisten ontologioiden kehityksessä.

Acknowledgements

As my journey to complete this thesis is finally coming to an end, I must express my deepest gratitude to those who I have had the honor to work with during my years at the university and to those who have supported my decision to pursue this career. I am sure I have forgotten to mention some of you by name, and I apologize for this oversight.

I would like to start by thanking my supervisors Filip and Tapio for your endless wisdom, guidance and encouragement. Your feedback and ideas have been invaluable to my research, but I am most grateful for your trust in my abilities to freely pursue the research directions I've deemed the most intriguing, and your patience while I have failed over and over again. I would also like to thank for the financial safety our research group has provided; something we take for granted, although this is not the case for many PhD candidates.

I would like to extend my appreciation to my fellow PhD students Suwisa, Farrokh, Juhani, Jenna, Li-Hsin, TurkuNLP members Hans, Jari, Sampo, Veronika, Aleksi, Antti V., Akseli, Aki and Samuel as well as my colleagues Antti A., Tapio P. and Martti from the Department of Computing. Thank you all for creating a supportive and fun work community. I miss all the long nights writing papers together until 4 a.m. and I hope our friendship will last a lifetime.

Special thanks to Hans, Sanna and the whole IKITIK consortium for welcoming me into the group with open arms. The clinical domain became the central part of this thesis and still remains my favourite application area of language technology. I hope some day I have the opportunity to return to these studies.

I would also like to thank Sofie and Sampo for your enthusiasm and passion for science. You are truly exemplar researchers and I will always look up to you.

Lasty, I must thank the reviewers of this thesis, Professor Øystein Nytrøn and Dr. Martin Duneld for their excellent criticism and feedback which I have tried to my best to incorporate into the final version of the manuscript. Further, I sincerely thank Professor Hanna Suominen for kindly agreeing to act as my opponent.

Turku, May 2024
Kai Hakala

List of original publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Moen*, H., Hakala*, K., Peltonen, L.-M., Suhonen, H., Ginter, F., Salakoski, T., and Salanterä, S. (2020b). Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. *Journal of the American Medical Informatics Association*, 27(1):81–88

II Moen*, H., Hakala*, K., Peltonen, L.-M., Matinolli, H.-M., Suhonen, H., Terho, K., Danielsson-Ojala, R., Valta, M., Ginter, F., Salakoski, T., et al. (2020a). Assisting nurses in care documentation: from automated sentence classification to coherent document structures with subject headings. *Journal of Biomedical Semantics*, 11(1):1–12

III Hakala*, K., Mehryary*, F., Moen, H., Kaewphan, S., Salakoski, T., and Ginter, F. (2017). Ensemble of convolutional neural networks for medicine intake recognition in Twitter. In *SMM4H@ AMIA*, pages 59–63

IV Moen*, H., Hakala*, K., Mehryary, F., Peltonen, L.-M., Salakoski, T., Ginter, F., and Salanterä, S. (2017). Detecting mentions of pain and acute confusion in Finnish clinical text. In *BioNLP 2017*, pages 365–372

V Hakala, K. and Pyysalo, S. (2019). Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61

VI Hakala*, K., Kaewphan*, S., Björne*, J., Mehryary, F., Moen, H., Tolvanen, M., Salakoski, T., and Ginter, F. (2020). Neural network and random forest models in protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*

* equal contribution

List of co-authored publications not included in the thesis

- Uronen, L., Salanterä, S., Hakala, K., Hartiala, J., and Moen, H. (2022). Combining supervised and unsupervised named entity recognition to detect psychosocial risk factors in occupational health checks. *International Journal of Medical Informatics*, 160:104695

- Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23

- Moen, H., Hakala, K., Peltonen, L.-M., Suhonen, H., Loukasmäki, P., Salakoski, T., Ginter, F., and Salanterä, S. (2018). Evaluation of a prototype system that automatically assigns subject headings to nursing narratives using recurrent neural network. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 94–100

- Kaewphan, S., Hakala, K., Miekka, N., Salakoski, T., and Ginter, F. (2018). Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. *Database*, 2018

- Sarker, A., Maksim, B., Jasper, F., Hakala, K., Svetlana, K., Mehryary, F., Han, S., Tran, T., Rios, A., Kavuluru, R., de Bruijn, B., Ginter, F., Mahata, D., Mohammad, S. M., Nenadic, G., and Gonzalez-Hernandez, G. (2018). Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*

- Mehryary, F., Hakala, K., Kaewphan, S., Björne, J., Salakoski, T., and Ginter, F. (2017). End-to-end system for bacteria habitat extraction. In *Proceedings of the 2017 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics

- Kaewphan, S., Mehryary, F., Hakala, K., Salakoski, T., and Ginter, F. (2017). TurkuNLP entry for interactive Bio-ID assignment. In *Proceedings of the BioCreative VI Workshop*, pages 32–35. Bethesda, MD, USA

- Hakala, K., Kaewphan, S., Salakoski, T., and Ginter, F. (2016). Syntactic analyses and named entity recognition for PubMed and PubMed Central —up-to-the-minute. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pages 102–107. Association for Computational Linguistics

- Mehryary, F., Kaewphan, S., Hakala, K., and Ginter, F. (2016). Filtering large-scale event collections using a combination of supervised and unsupervised learning for event trigger classification. *Journal of Biomedical Semantics*, 7(1):1–13

- Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):1–19

- Hakala, K. (2015). UTU: Adapting biomedical event extraction system to disorder attribute detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 375–379. Association for Computational Linguistics

- Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2015). Application of the EVEX resource to event extraction and network construction: Shared Task entry and result analysis. *BMC Bioinformatics*, 16(Suppl 16):S3

- Pyysalo, S., Campos, J., Cejuela, J. M., Ginter, F., Hakala, K., Li, C., Stenetorp, P., and Jensen, L. J. (2015). Sharing annotations better: RESTful Open Annotation. In *Proceedings of ACL'15: Demonstrations*

- Mehryary, F., Kaewphan, S., Hakala, K., and Ginter, F. (2014). Eliminating incorrect events from large-scale event networks by trigger word clustering and pruning. In *Proceedings of SMBM'14*, pages 75–79

- Kaewphan, S., Hakala, K., and Ginter, F. (2014). UTU: Disease mention recognition and normalization with CRFs and vector space representations. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 807–811. Association for Computational Linguistics and Dublin City University

- Van Landeghem, S., Björne, J., Wei, C.-H., Hakala, K., Pyysalo, S., Ananiadou, S., Kao, H.-Y., Lu, Z., Salakoski, T., Van de Peer, Y., et al. (2013).

Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814

- Hakala, K., Mehryary, F., Kaewphan, S., and Ginter, F. (2013a). Hypothesis generation in large-scale event networks. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM'13)*, pages 19–28

- Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2013b). EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop (BioNLP-ST'13)*, pages 26–34

- Van Landeghem, S., Hakala, K., Rönnqvist, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics, special issue Literature-Mining Solutions for Life Science Research*

- Hakala, K., Van Landeghem, S., Kaewphan, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). CyEVEX: Literature-scale network integration and visualization through cytoscape. In *Proceedings of SMBM'12, Zurich, Switzerland*, pages 91–96

Abbreviations

BiLSTM Bidirectional Long Short-Term Memory

BLAST Basic Local Alignment Search Tool

BoW Bag-of-words

BPE Byte Pair Encoding

CNN Convolutional Neural Network

CRF Conditional Random Field

CV Cross-validation

EHR Electronic Health Record

EMR Electronic Medical Record

FinCC Finnish Care Classification

FNN Feedforward Neural Network

GLUE General Language Understanding Evaluation

GO Gene Ontology

GPT Generative Pre-trained Transformer

HT Homology Transfer

IOB Inside, Outside, Beginning

LSTM Long Short-Term Memory

ML Machine Learning

MLM Masked Language Modeling

MRR Mean Reciprocal Rank

NER Named Entity Recognition

NLP Natural Language Processing

OOV Out-of-vocabulary

POS Part-of-speech

ReLU Rectified Linear Unit

RF Random Forest

RNN Recurrent neural network

SC Sentence Classification

SGD Stochastic Gradient Descent

SVC Support Vector Classifier

SVD Singular Value Decomposition

SVM Support Vector Machine

TFIDF Term Frequency–Inverse Document Frequency

Contents

1 Introduction

1.1 Natural language processing and text mining

Natural language processing (NLP) is a subfield of computer science and linguistics focusing on the analysis, understanding, manipulation and generation of human language. As the emphasis of NLP is often on algorithm development for automatically processing large quantities of natural language data and improving the interaction between humans and machines, it is strongly influenced by the scientific disciplines of *linguistics* and *machine learning*, and is often indistinguishable from *computational linguistics* in terms of methodology. Linguistics and the *philosophy of language* establish the theoretical foundations for modelling human language in NLP, although these theories are often blatantly streamlined in practice. As the complexity of human language limits the feasibility of rule-based solutions, it has driven computer scientists to rely on machine learning methods instead. Identical statistical and machine learning models are often developed and used in both natural language processing and computational linguistics although the applications may differ greatly. The more engineering-oriented audiences prefer the term *language technology* instead of natural language processing.

During the last decade, the best-performing NLP methods have relied more and more on *representation learning* (Bengio et al., 2013), the approach of forming latent representations of the patterns in the underlying data, instead of emphasizing manually crafted features. This in turn has shifted the field towards studying more suitable, and often more complex, machine learning models that empirically demonstrate strong performance, but on the other hand are often challenging to interpret and analyze due to the nonlinear nature of the latent representations (Wiegreffe and Pinter, 2019; Wallace et al., 2019). Although the mathematical definitions of the used models vary greatly, many of them are often grouped under the umbrella term *neural networks* and share some similarities, e.g. learning algorithms based on stochastic gradient descent.

The increasing complexity of these neural models has created a new challenge as the task-specific training datasets available are often too limited for learning generalizable latent representations, leading to the exploration of *transfer and multitask learning*, where the models are trained on multiple tasks (Torrey and Shavlik, 2010; Caruana, 1997). The expectation is that the latent representations of human language learned from one task are beneficial for other tasks, thus reducing the amount of task-

specific data required. *Pretraining*, i.e. the method of initially training a model with a universal language modelling task and subsequently learning the actual task at hand, has become the prevalent approach for utilizing such learning strategies in NLP.

Although natural language processing covers the analysis of human language on various levels and from different aspects, *text mining* specifically refers to the application of NLP methods to extract and derive semantic-level information from large quantities of textual data and may include tasks such as *text classification*, *named entity recognition* and *relation extraction*. The task of text classification aims at grouping texts, e.g. sentences or documents, into predefined categories, often based on their meaning or topic (Kowsari et al., 2019). Named entity recognition on the other hand focuses on identifying mentions of entities that are relevant for a given task, e.g. drug names or locations, and relation extraction seeks to determine how the given entities are connected (Mikheev et al., 1999; Bach and Badaskar, 2007). Although these tasks may sound simple, they form the basis of automatically extracting and connecting small nuggets of information necessary for forming new hypotheses and facts from the abundance of data.

It must be pointed out that although recent neural methods for text mining show promising results even on tasks which seemingly require high-level semantic understanding of language and common sense reasoning, it is unclear how these models internally represent the information they have learned. Moreover, there are some indications that these models rely heavily on fallacious statistical cues (Niven and Kao, 2019) and may not actually learn any meaningful representations of the semantic characteristics of the targeted texts, that would provide further knowledge of the given task. Whether this is an indication of poorly designed problem definitions and evaluation metrics or an immature technology remains to be seen.

1.2 Textual data in biomedical research and clinical work

Global yearly spending on life science research sums up to hundreds of billions of US dollars (Chalmers and Glasziou, 2009) leading to a colossal amount of available research reports in the biomedical field. At the time of writing this thesis, PubMed contains over 30 million biomedical research articles and over 3000 new articles are published in peer-reviewed journals daily (Lee et al., 2019). Moreover, the amount of annual publications has been increasing for the past 60 years, annual growth doubling every thirteen years (Larsen and Von Ins, 2010). At the same time, the quality of scientific reporting has been strongly criticized and claims such as over 50% of biomedical research reports being unusable due to poor quality or missing information have been made (Glasziou et al., 2014). It is no surprise that these issues are thus habitually mentioned as the main culprits for the growing demand for NLP tools suitable for biomedical literature as new biomedical discoveries and insights are lost in the overwhelming amount of largely unusable literature.

In the clinical domain the documentation has a more immediate role as much of the administered care and the observations of the condition of the patient are written in free text narratives on regular basis, often multiple times per day during a hospital stay. Although this information can accumulate to extensive medical histories, it should be readily available and easily accessible for efficient clinical decision-making. However, in practice these records tend to cause data management and information overload issues (Hall and Walton, 2004). Information systems providing only the relevant information for the current decision to be made are much sought after in studies focusing on medical information overload (Hall and Walton, 2004; Hanka and Fuka, 2000; Hunt and Newman, 1997).

Another concern, particularly in the clinical domain, is the efficiency of producing high quality documentation, as the resources spent on care documentation could alternatively be allocated on actual patient care. Prior studies suggest that documentation can consume over 35% of medical-surgical nurses' working time (Hendrich et al., 2008), more than what is spent on actual patient care activities (Momenipour and Pennathur, 2019). Alarmingly it also seems that the emergence of electronic health and medical record (EHR/EMR) systems has not improved the efficiency of clinical care documentation (Yee et al., 2012), but may actually decrease care efficiency (Schenk et al., 2018). Whereas well-thought EHR skills training can improve the situation (Robinson and Kersey, 2018), it is clear that natural language processing methods are needed for producing and utilizing clinical documentation effectively. As the secondary use of clinical data has been recognized as essential for healthcare management, public health management and clinical research (Meystre et al., 2017), additional opportunities and challenges for clinical text mining arise.

A third distinctly different, yet exceedingly important source of textual data used in biomedical studies is social media, with various applications e.g. in public health monitoring (Paul and Dredze, 2011) and pharmacovigilance (Sarker et al., 2015). Although online users are concerned about how their personal information is used, social media platforms have been widely adopted for health information exchange (Lin and Chang, 2018) enabling the application of NLP methods for large-scale statistical analyses. However, using such user-generated content in medical studies has raised concerns, mostly due to the quality and reliability of the data (Moorhead et al., 2013), making the automated processing of such data more challenging, yet also bringing forth the need for more sophisticated natural language processing methods in this domain.

Although here we have focused on textual data, similar issues are seen with high-throughput nucleic acid sequencing in life sciences as the number and diversity of sequenced genomes is growing at an unprecedented rate (Goodwin et al., 2016), yet characterising the functional properties of the newly discovered nucleic and amino acid sequences remains a challenge due to the costly and time-consuming experiments (Zhou et al., 2019). Similarly to NLP methods being used in extracting valu-

able information from the biomedical literature, computational methods in life sciences, e.g. in the form of protein function prediction, have sparked research interest to keep up with the efficiency of the high-throughput sequencing methods.

1.3 Research objectives

This thesis focuses on clinical and biomedical text mining, in particular on text classification and named entity recognition, with the aim of advancing the development and evaluation of natural language processing methods to the point of being beneficial in real-world applications. As the recent trends in NLP have strongly favored neural network based models, the main weight is on developing text mining tools based on a range of neural architectures proposed during the past years. The different variants of neural models are benchmarked against each other as well as against more traditional machine learning methods as a natural question with the emerging technologies is whether they provide clear benefits over the previous generation of NLP techniques.

Although the need for natural language processing methods in the biomedical field has been recognized, the adoption of these methods in biomedical research and clinical work has been few and far between (Zheng et al., 2015). Poor usability, maintainability and integration into real workflows are often mentioned as the major issues (Chapman et al., 2011). From the adoption perspective, this thesis tries to assess whether the current state-of-the-art methods are good enough to be applied in practice by taking into account human-level performance in the studied applications, a question often left unanswered in natural language processing evaluations. Whereas human-level performance may not always be a good indicator of the applicability of a method, in this thesis the primary focus is on a currently manual task, which could be automated if the performance of the domain experts could be achieved.

Whereas clinical text mining has been widely studied, these studies tend to be heavily biased, with most experiments conducted on pathology and radiology reports, whereas e.g. nursing notes have been mostly disregarded (Mujtaba et al., 2019). Moreover, the vast majority of the studies address only English data and although modern NLP methods are language-agnostic, it is unclear whether the findings of these studies are directly transferable to other languages. Thus, in this thesis the main focus has been placed on Finnish nursing notes, on which both text classification and named entity recognition tasks are discussed. Text classification supporting biomedical research is further expanded to English social media texts, whereas named entity recognition is also studied in the context of Spanish clinical case reports in medical publications, making it possible to draw conclusions from three distinct text sources and three different languages. As the resources needed for pretraining the recent NLP models have become unobtainable for most researchers

and developers this setting also enables the study of multilingual and cross-domain knowledge transfer.

As briefly mentioned in Section 1.2 it is unclear how neural models represent information and whether the learned representations are only beneficial for solving the task at hand or whether they have the potential for secondary use cases. In this thesis one of the goals is to analyze the label representations the neural text classification methods form in an attempt to assess whether they reveal previously unknown peculiarities in clinical language use.

The discussion is broadened to protein function prediction, which in this context is considered analogous to text classification due to the sequential nature of both of these data sources.

To summarize, the main objectives can be described as:

1. Develop and evaluate neural approaches for clinical and biomedical text classification and named entity recognition.
2. Address the question whether the evaluated methods are good enough to be deployed for actual use in clinical care.
3. Study neural models in multilingual and cross-domain settings.
4. Explore methods for studying the learned representations of neural models and their applicability to secondary uses in analyzing clinical care documentation.
5. Explore the possibility of treating amino acid sequences analogously to text classification in the scope of automated functional annotation.

1.4 Contributions of the work

Notice that this thesis is a compilation of studies mostly conducted between years 2016 and 2019 in a technological turmoil, as more sophisticated pretraining methods revolutionized the field in 2018 (Devlin et al., 2019; Radford et al., 2018) (see Section 2.3.1). Thus the methods used in the earlier studies summarized in this thesis are already to a large extent obsolete, although they represent the state-of-the-art of their time. Consequently, at the time of writing this summary, many of the used methods should no longer be used as starting points for future work. In this thesis I have thus tried to emphasize more the general observations and findings discovered in our studies rather than the technological details of our methods.

The main contributions of the papers included in this thesis are:

- **Paper I:** *Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods* applies and evaluates a selection of widely used text classification methods for a nursing note classification task with weakly labelled

training data. The study suggests that current state-of-the-art methods are on a par with domain experts, permitting the usage of such methods in real-world clinical applications.

- **Paper II:** *Assisting nurses in care documentation: from automated sentence classification to coherent document structures with subject headings* demonstrates that the text classification methods from Paper I can be used to automatically restructure free clinical narratives into semi-structured nursing notes and that the internal knowledge representations that the neural classification models form are beneficial in studying clinical documentation practices and problematic ontology definitions.

- **Paper III:** *Ensemble of convolutional neural networks for medicine intake recognition in Twitter* presents a text classification model which combines dense and bag-of-words style feature representations for social media data and discusses the benefits of a model ensemble in reducing variance caused by random model initialization.

- **Paper IV:** *Detecting mentions of pain and acute confusion in Finnish clinical text* describes experiments on a named entity recognition task in a low-resource setting and discusses the relationship of named entity recognition and text classification tasks. The study shows that reducing a named entity recognition task into a text classification task may not actually simplify the task or lead to a better performance.

- **Paper V:** *Biomedical named entity recognition with multilingual BERT* demonstrates that simple cross- and multilingual approaches with readily available pretrained models and tools can lead to strong performance, even outperforming methods specifically tailored for the given language and domain.

- **Paper VI:** *Neural network and random forest models in protein function prediction* introduces an ensemble method of neural networks and random forests for protein function prediction utilizing a wide range of numerical features as well as the plain amino acid sequences. In addition to the demonstrated strong performance of the full system, the study shows that functional properties of proteins can be directly learned from the amino acid sequences with neural models analogous to text classification methods.

1.5 Structure of the thesis

The thesis consists of four chapters, the first being this introductory section describing the research objectives and the general background of the work. In Chapter 2 the key components and concepts of neural natural language processing are presented.

These form the machine learning foundations for all the studies included in this thesis. The papers included in this thesis are summarized in Chapter 3. This chapter is divided into three distinct sections, the first focusing on text classification, the second on named entity recognition and the third on protein function prediction. The papers are also available as reprints in the second part of this thesis. The Chapter 4 provides a retrospective inspection of the studies, their contributions, their mutual relations and connection to subsequent studies conducted by other researchers. Plausible future research directions are also discussed and recommended in the last chapter.

2 Foundations: Neural models in natural language processing

In this chapter we go through the standard components and training methods used in neural networks, particularly in NLP applications. As many of these approaches are utilized in the studies incorporated into this thesis, the general ideas behind the models and their fundamental differences are collectively described in this chapter, whereas the latter chapters describe the specific models used in the studies only on a higher level and assume that the readers are familiar with the concepts mentioned in this chapter. The main focus in this chapter is on the structural differences of commonly used neural architectures and their components as well as on the ways these models can be trained with and without task-specific training data.

2.1 Components of neural architectures

2.1.1 Fully connected layers, latent word representations and sequential modelling

The simplest form of neural networks comprises of a linear mapping of an input vector to an output scalar. The output value is calculated as a weighted sum of the input values (dot product of the input vector and the weight vector) and often normalized with a selected activation function. Learning with such a model was introduced in the 1950s as the (single-layer) Perceptron, although equivalent formulations in the form of linear regression were studied already in the early 19th century (Schmidhuber, 2015). The output can be generalized from a scalar value to a vector by using multiple Perceptron nodes in parallel, each learning its own set of weights, but sharing the same input vector. In this thesis we refer to this neural architecture as a fully connected layer.

Fully connected layers are an essential building block in modelling language: they tend to form at least the first and last layers of commonly used neural architectures in NLP. A natural way of representing words is a vector where each element represents a unique token in the vocabulary. This form is used in traditional bag-of-words feature representations, but can also be used to represent single words: the element corresponding to the given word is set to one, whereas other elements are set to zero (one-hot encoding). If we are to model e.g. the semantic similarities of

words, these word representations however cause a problem: all one-hot encoded word vectors are orthogonal with each other and the representations are not able to model whether e.g. one pair of words is semantically more similar than another pair. Thus, using a fully connected layer to project the one-hot encoded word vectors to latent representations, often called word embeddings, has become the standard approach for modelling individual words in NLP. These representations tend to be less than 1000 dimensional, whereas the one-hot encoded vectors reflect the vocabulary size and are generally orders of magnitude larger. Notice that instead of explicitly modelling the word embeddings as a fully connected layer, they can also be seen as a table look-up where the index of the non-zero value in the one-hot encoded word vector refers to the corresponding set of weights stored in the table. This also brings forth the fact that the projection of a one-hot encoded vector is equivalent to the learned weights corresponding to the non-zero value in the fully connected layer (as other weights are multiplied by the zero values in the dot products) if linear activation function is used. How these latent representations can be trained to model meaningful word relations without any manually labelled data is discussed in Section 2.3.

Whereas traditional bag-of-words feature representations lose word order information, this information can be preserved in neural models by forming the input as a sequence of one-hot encoded word representations and consecutively as a sequence of word embeddings. Thus, the input features form a matrix instead of a single vector. Such a sequential representation was effectively present already in the early neural n-gram language models (Bengio et al., 2003), although in practice the input was formed as a concatenation of the word embedding sequence.

Notes on tokenization

Although often referred to as word embeddings it is good to point out that similar representations can be used for various segments of text, often called *tokens*. A challenge with models using word-level representations is that the vocabulary size grows to the level of hundreds of thousands or millions of unique words, each represented with a distinct vector. Often many of the rare words have to be replaced with out-of-vocabulary (OOV) tokens to reduce the memory consumption of the models and due to the difficulty of learning good representations for rare words (Bojanowski et al., 2017).

At the other extreme are character-level models, where the embedding set is limited to the set of unique letters, requiring orders of magnitude less memory. In such models the semantic representation of the input text is more dependent on the upper layers of the network as the non-contextualized embeddings only represent single characters. Albeit representing a language on character-level seems challenging, strong performance has been demonstrated with character-based models (Akbik

et al., 2018). As the character-level embeddings are in principle able to represent all possible words of a language, they are often combined with word-level representations to fill in the gaps where the model is unable to learn strong word-level representations directly or for situations where a previously unseen word is observed (Bojanowski et al., 2017; Ma and Hovy, 2016).

The challenge with character-level models is that deriving the meaning of a word purely from a sequence of characters can be very demanding task to learn, whereas with word-level embeddings some semantic information can be encoded directly into the word vectors. To combine the strengths of both extremes, most recent models use varying size subword tokens. In these methods the vocabulary is set to a fixed size in advance and the tokenization is done in a way that preserves common words intact and splits rarer words into common subwords based on word occurrence statistics gathered from a corpus. These subwords may not follow any linguistically meaningful units, but can be mere common character n-grams. The full character set is also included in the vocabulary and thus the models are able to form representations for all possible words, but at the same time the most common words can be directly represented with a single vector. The exact way of defining the words and subwords that are included in the vocabulary differ slightly between different tokenizers (Sennrich et al., 2016; Kudo and Richardson, 2018). For instance in the byte-pair encoding tokenization (BPE) the vocabulary is first populated with all unique characters and subsequently the most frequently appearing token pair, starting from character pairs, is repeatedly added to the vocabulary and the merge operation for joining the given token pair is recorded until the desired vocabulary size is reached (Sennrich et al., 2016). As a result, the vocabulary will contain common words, common subwords and the full character set. During tokenization OOV words can be similarly split into subword tokens by starting from the plain character sequence and repeatedly applying merge operations until no possible operations are available anymore. Notice that BPE tokenization requires the text to be pre-tokenized into individual words, whereas other methods such as SentencePiece (Kudo and Richardson, 2018) do not have this limitation.

2.1.2 Convolutional neural networks

In contrast to fully connected layers, convolutional neural networks (CNN) rely on filters, which are applied on a local region of the input. Whereas CNNs were originally popularized by their success in computer vision (Krizhevsky et al., 2012; Szegedy et al., 2016), they have seen broad adoption in natural language processing (Ma and Hovy, 2016; Zeng et al., 2014; Zhang et al., 2015; Hughes et al., 2017). Kim (2014) suggested one of the earliest architectures incorporating word embeddings and shallow CNNs, with a single convolutional layer. Unlike 2-dimensional CNNs in computer vision, this model uses a collection of convolutional filters which

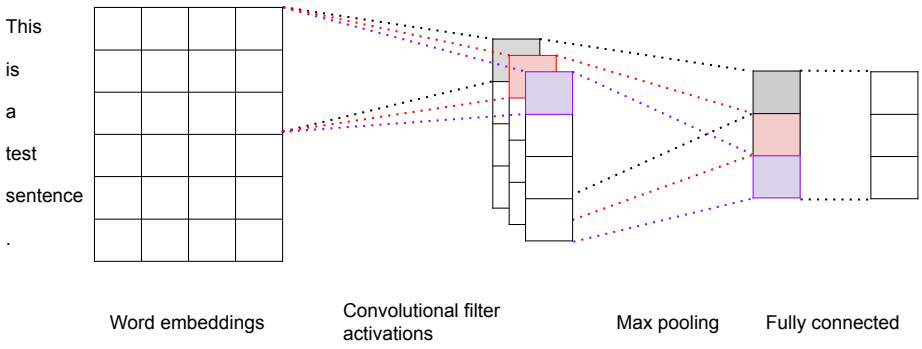

Figure 1. An example of a convolutional text classification architecture. Each word is represented with a word embedding and a collection of three convolutional filters (of width 3) is applied in a sliding window across the word sequence. Each filter activation sequence is further max pooled and the one-dimensional encoding is subsequently fed to a fully connected output layer. A filter consists of a weight matrix identical in shape to the targeted input window (3 x 4 in this example).

are applied on a local region (receptive field) of the input word embedding sequence in a sliding window, solely across the word sequence dimension. As the output, each filter produces a single activation scalar for each timestep in the input by summing the element-wise multiplication of the input window and the learnable filter weights, thus each activation value describes the fitness of the filter and the targeted word subsequence. The output can be further scaled with any activation function, often rectified linear activation (Nair and Hinton, 2010; Xu et al., 2015).

Kim (2014) use such modelling for text classification, but as the convolutional layer produces an output array of order 2, they reduce the dimensionality to a one-dimensional array by selecting the maximum values for each filter (max pooling). This encoding is subsequently fed as an input to a fully connected output layer. A visual representation of this type of a network is depicted in Figure 1. In this thesis variants of this architecture as well as other usages of convolutional layers are described in Sections 3.1.2, 3.1.3 and 3.3.

2.1.3 Recurrent neural networks

Whereas CNNs are able to model word order and word relations within their receptive fields, the activations of different timesteps are independent of each other. Although this property is desirable from computational perspective as the activations

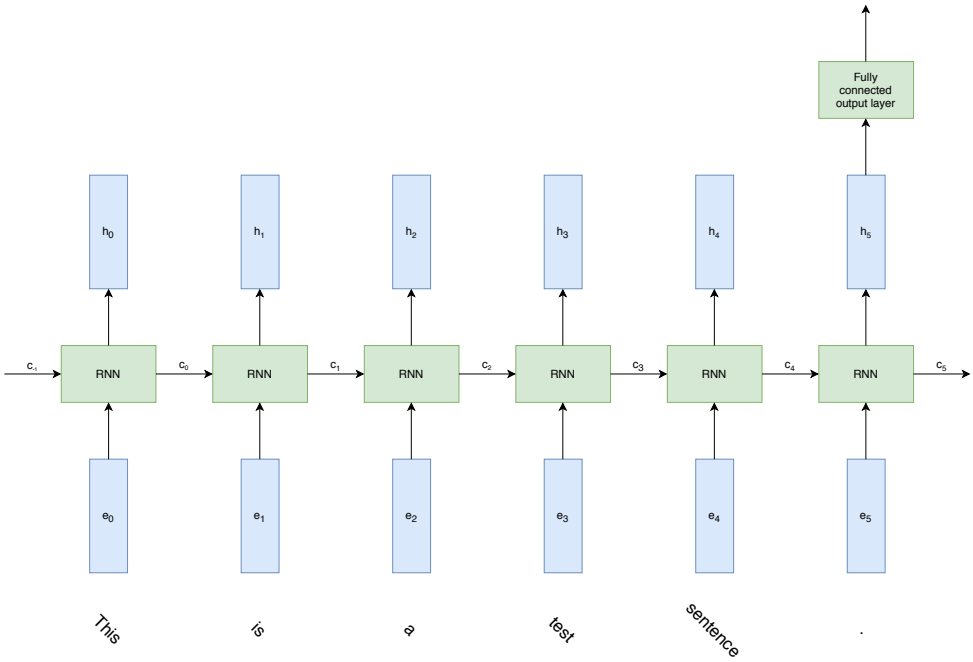

Figure 2. An unrolled example of an RNN text classifier architecture. Each output is produced based on the current input embedding (e_t) and the recurrent connection (c_t). All RNN nodes share the same weights. Notice that in vanilla RNNs vectors h_t and c_t are identical (excluding the bias term), whereas in an LSTM an additional cell state vector is passed in the recurrent connection. In text classification the RNN output vector for the last timestep (here h_5) is often used as an encoding of the whole input sequence and fed to a fully connected output layer, while the other output vectors are ignored.

can be calculated concurrently, it limits CNNs capability to model long distance relations in the input word sequence. A single convolutional layer is able to model the relations of the whole input sequence if and only if the width of the receptive field is equal to the input sequence length, in which case the convolutional layer acts as a fully connected layer over the concatenated input word embeddings.

An alternative group of neural models called recurrent neural networks (RNNs) model the sequential dependencies in the input by conditioning the output of each timestep on an output of the previous timestep (Sutskever et al., 2011). In the simplest form an RNN can be a fully connected layer with the input of the current timestep and the output of a previous timestep concatenated as the actual input to the layer (often referred as a vanilla RNN) (Sutskever et al., 2011). An example of an RNN is shown in Figure 2.

Notice that to compute the activation for a given timestep, all previous timesteps have to be first computed. Similarly, during training the error signal is backpropagated through time, i.e. through the unrolled computational graph. In a vanilla RNN,

the propagated error over timesteps depends exponentially on the layer weights, which often leads to either vanishing or exploding gradients (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997). As a result, such RNNs tend to be difficult to train to model long distance dependencies. To overcome this, Hochreiter and Schmidhuber (1997) suggested an architecture called long short-term memory (LSTM) networks, which along with its variants has since become the standard RNN approach in language technology. In essence, LSTMs incorporate an additional output (often called a cell state) which is passed over the timesteps in addition to the actual activations. The actual output and the cell state are decoupled with a set of learnable gate functions leading to non-exponential dependencies between the error propagation and layer weights, thus mitigating the vanishing and exploding gradient issue.

In text classification the RNN outputs for each timestep can be pooled in a similar fashion as with CNNs, but the more common approach is to only use the output of the last timestep, while ignoring the activations for all other timesteps (Lee and Dernoncourt, 2016) (see Figure 2). In this approach the RNN is directly encoding a semantic representation of the input text into a single vector.

The standard RNN models only condition the output of the given timestep on the previous timestep, resulting in word representations, which only take into account the left context of the target word. In some NLP tasks better results can be achieved with a simple trick of adding an independent recurrent layer with inverse time direction and concatenating the outputs from the original and the inversed RNN layers. The outcome is often called a bidirectional RNN (Schuster and Paliwal, 1997). Notice that a single bidirectional RNN layer models both left and right contexts independently and is thus not able to combine the information from both directions. Multiple RNN layers should be stacked for incorporating information from both contexts.

In this thesis RNN-based models are explored in Sections 3.1.2 and 3.2.2.

2.1.4 Attention and Transformers

In the previous sections we have described how CNNs and RNNs are commonly used for modelling sequential data and how these models either form representations for each timestep in the input sequences or for the whole sequence. These architectures are suitable for instance for text classification and sequence labelling (including named entity recognition), which are the main topics of this thesis. However, in many tasks a model able to produce arbitrarily long output sequences is needed. The most intuitive example of such task is probably neural machine translation, where a word sequence in the source language is firstly encoded with a neural model and subsequently an autoregressive decoder predicts the corresponding output word sequence in the target language conditioned on the encoded source sequence (Sutskever et al., 2014). Such neural encoder-decoder models are also called sequence-to-sequence models.

As the only knowledge transferred between the encoder and decoder is a single vector encoding of the input sequence, e.g. the last timestep activations of an LSTM layer, an information bottleneck occurs between these model components. Bahdanau et al. (2014) suggested a solution to this problem, by providing the decoder an access to the weighted average across all encoder timesteps (e.g. encoder LSTM activations for all input steps). In their model the weighting of the encoder outputs was conditioned on the current state of the decoder, i.e. the decoder had the opportunity to learn which words in the input sequence it deemed important for the next output word to be predicted. Bahdanau et al. (2014) called this conditioned weighting an *alignment model*, but the name *attention* (which they also used to describe the model) quickly became the more popular term for this concept.

On an abstract level attention can be described with three sets of vectors: *keys*, *values* and *queries*. For each timestep in the input sequence a *key*, a *query* and a*value* are formed with fully connected layers. These can simply be linear projections of the input vector. A *query* vector is compared against all *key* vectors to measure their compatibility, e.g. by simply taking a dot product, and these compatibilities are further normalized with a softmax activation to form a probability distribution of the key vectors. A weighted sum of the value vectors is subsequently produced using the normalized compatibility values as the weights to form an output vector (see Figure 3). Notice that *query*, *key*, *value* and output vectors have identical dimensionality and instead of vectors, the computations can also be represented with matrices. The attention output for each timestep can be calculated independently of other timestep outputs unlike in RNNs, but the formed encoding contains both preceding and following contexts.

Later studies demonstrated that attention was not only beneficial as a supporting element in sequence-to-sequence models, but in fact powerful enough to encode sequential data in itself. Vaswani et al. (2017) suggested an architecture where the encoding for each input step is formed as a weighted average of all inputs, conditioned on the given timestep, i.e. *self-attention* (Cheng et al., 2016; Parikh et al., 2016). As the same weighting is applied on each dimension of the input vectors, i.e. each input vector has a single weight value, a single attention mechanism is not able to attend on different aspects of the inputs. Vaswani et al. (2017) solved this by using multiple independent attentions over the same input sequence (multi-head self-attention). The output vectors from all attention heads are concatenated and projected with a fully connected layer to the original dimensionality of the input vectors, which permits stacking arbitrary amount of such blocks. Additional fully connected layers and residual connections are included in the actual implementation. Vaswani et al. (2017) call this architecture the *Transformer* and show that both encoder and decoder components can be built using them, without any recurrent or convolutional layers.

Notice that although self-attention is able to include information from both pre-

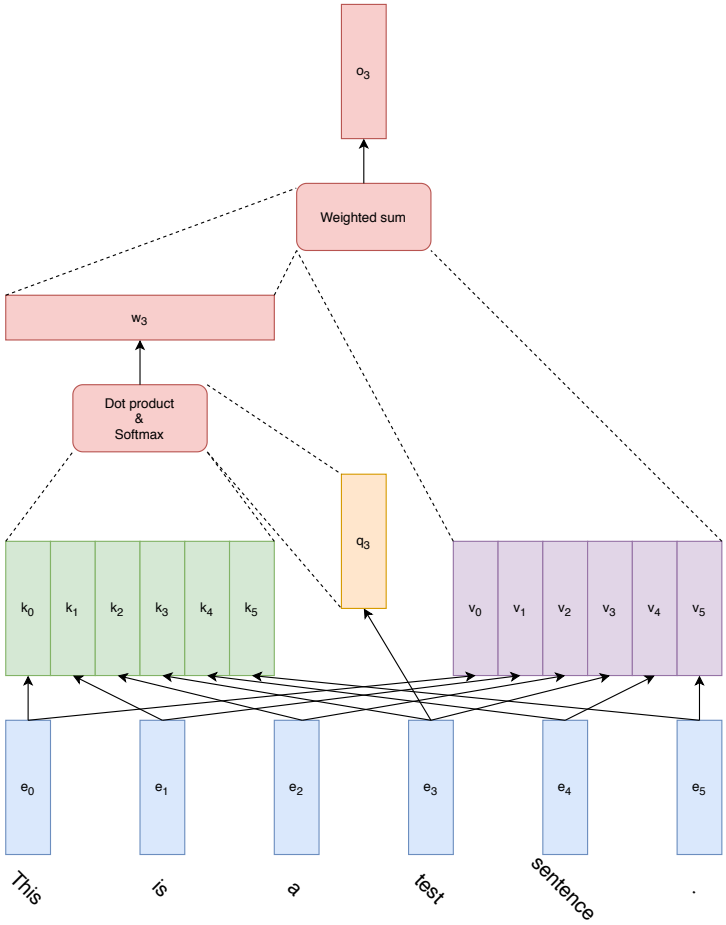

Figure 3. Computation of a single self-attention head for a given input timestep (e_3 in this case). Notice that *key* and *value* vectors are formed even for the targeted timestep, i.e. the model attends on all words in the sequence, not only on the surrounding words. For the targeted word also a *query* vector is formed. Dot product between the *query* and *key* vectors is computed to form a relative weighting and subsequently a weighted sum, which is the output of the attention.

vious and following words, the weighted sum does not take into account word order. In this sense self-attention is a weaker model than CNNs and RNNs, although it also enables the model to represent long distance word relations with a constant computational complexity, unlike RNNs. To incorporate some sequential information to the model, the standard approach is to use positional embeddings (Vaswani et al., 2017). For each position in the input sequence a vector representation is constructed and subsequently joined with the actual word embedding, usually by simply summing the positional and word embeddings. In practice these positional embeddings can either be learned or precalculated, e.g. based on trigonometric functions (Vaswani

et al., 2017), although learning such embeddings reduces the model's generalization ability to sequences which are longer than the sequences in the original training data.

In this thesis the Transformer architecture is explored for a named entity recognition task in Section 3.2.3.

2.1.5 Conditional random fields

The previously described neural architectures are all able to produce token-wise semantic representations of the underlying text. In token-level classification tasks, such as named entity recognition (see Section 3.2), these representations are often used as an input for a decision layer doing the final prediction. However, in such sequential classification tasks, the labels of neighbouring tokens are usually strongly dependent on each other. As an example, in part-of-speech tagging for Finnish, it would be very unlikely that two coordinating conjunctions appear successively, whereas a higher likelihood can be expected for a *noun* label, if the previous label is an *adjective*. Using a fully connected token-wise output layer does not explicitly take these dependencies into account, although the underlying token representations might implicitly model such relations and the recent Transformer-based models indeed do (Devlin et al., 2019; Hakala and Pyysalo, 2019).

A popular approach for explicitly modelling these label dependencies is the family of conditional random field models (CRFs) (Lafferty et al., 2001), in practice the linear chain variants. Similarly to hidden Markov models, CRFs can be used to model the transition likelihoods across states, i.e. in our case the likelihoods of a subsequent label, given the previous label. In essence, CRFs model the conditional probability of a label sequence given the input feature sequence and in fact, both hidden Markov models and multinomial logistic regression models are special cases of CRFs (Sutton et al., 2012). Finding the optimal label sequence given a sequence of feature vectors can be done by enumerating all possible label sequences, which can be optimized with dynamic programming (Sutton et al., 2012).

As CRFs condition the output on a sequence of input features, they can be utilized in neural networks by considering the token-wise encoder outputs as the input features for the CRF layer. However, the way the CRF model conditions the outputs on the input features varies across implementations. E.g. Lample et al. (2016) implement a version where the likelihood of a label sequence is the sum of learnable transition scores added to the sum of label-specific scores conditioned on the underlying neural model. A contemporaneous study by Ma and Hovy (2016) on the other hand condition the transition scores on the input vector and do not incorporate a separate label-specific term in the scoring function [1].

[1] The approach by Lample et al. (2016) gathered more traction and has been implemented e.g. in TensorFlow, but in our unpublished experiments on various biomedical NER datasets, the variant by Ma and Hovy (2016) showed stronger results.

2.2 Training and regularization

The training of neural networks is generally accomplished by defining a differentiable loss function which measures the fitness of the model on the given task. This loss function is then optimized with stochastic gradient-based methods. These methods measure the loss function for a subsample of the training data, calculate the first-order partial derivatives of the function w.r.t. the model parameters and update these parameters proportionally to the gradients. In practice the loss functions are constructed as the negation of the model fitness and thus minimized, and the model parameters are adjusted based on the negative of the gradients, often called stochastic (or mini-batch) gradient descent (SGD) (Bottou, 1998). The process is repeated for different subsamples, thus minimizing the expected value of the loss function.

A variant of the standard SGD, called Adam (Kingma and Ba, 2015) has become one of the most widely used optimization algorithms in NLP research. Adam extends the normal SGD update rule with two additional concepts: the exponential moving average of the partial derivatives and the squared partial derivatives. These provide estimates for the mean and uncentered variance of the gradient. In Adam the parameter updates are essentially scaled with the ratio of these two, which results in higher effective step sizes when the uncentered variance is low in comparison to the mean and small step sizes when the variance is large. Thus, the ratio acts as a measure of reliability of the given gradient update (the authors refer to it as the signal-to-noise ratio). A practical concern, not so often discussed in research papers, is that for each model parameter two additional variables have to be tracked, which greatly increases the memory overhead of Adam compared to vanilla SGD.

Due to the high capacity of the common neural architectures and the efficiency of the gradient based optimization methods neural models often suffer from overfitting and require strong regularization techniques. In addition to the common L2 regularization (or the equivalent weight decay), variants of dropout (Srivastava et al., 2014) and batch normalization (Ioffe and Szegedy, 2015) are often used. For instance in the previously discussed *Transformer* architecture (see Section 2.1.4) weight decay, dropout and layer normalization (Ba et al., 2016) are all used jointly.

In dropout all units in a layer have a predefined probability of being removed for a single training pass. This enforces the model to rely on all parameters of the model as on some iterations the most influential features or parameters may end up being disconnected from the network. The probability of dropping out unit is usually a hyperparameter to be adjusted correctly. During inference all units are again activated[2], but to compensate for the higher values of the learned weights, they are down-scaled based on the chosen dropout probability. The intuition the authors provide for dropout is that the whole scaled-down network approximates an ensemble

[2] Sometimes dropout is applied even during inference to produce non-deterministic results (Shen et al., 2018).

of the sub-networks with dropped out units, leading to similar performance gains that is often seen in model combinations (Srivastava et al., 2014). Many variants of the dropout approach exist and e.g. instead of dropping out single units from the network it is possible to drop the whole word embeddings in the input sequence, by either replacing them with zero embeddings or an *unknown word* token (Iyyer et al., 2015).

A notable flaw in gradient descent algorithms is that during each training step model parameters are updated independently, i.e. under the assumption that the rest of the network remains unchanged, but still simultaneously. Thus a layer has to adapt to a constantly changing distribution of input values from a previous layer during training, a phenomenon called *internal covariate shift* (Ioffe and Szegedy, 2015). In practice, this issue leads to poor convergence, sensitive learning rates and long training times. Ioffe and Szegedy (2015) solve this issue by normalizing the output values of layers to zero mean and unit variance, thus preserving the input value distribution of the next layer. The normalization statistics (original mean and variance) are updated after every training step and as calculating such statistics from the whole training data is computationally expensive they simplify the approach by estimating the statistics from a single batch, leading to the method being called *batch normalization*. Another simplification in the approach is that each targeted input feature is normalized independently of other input features.

Ioffe and Szegedy (2015) apply the normalization before the activation function of a layer and therefore the normalization may for instance force the values on the linear region of a sigmoid function, reducing the expressive power of the model. To overcome this issue, they introduce learnable scaling and bias variables for each normalized feature, allowing the optimization method to adjust the strength of the normalization or disable it if needed.

Although the original motivation of *batch normalization* is to stabilize and accelerate model convergence, Ioffe and Szegedy (2015) empirically show that it also acts as a form of regularization and reduces the need for dropout and L2 regularization. A weakness of the method is that for small batch sizes the mean and variance estimates may be inaccurate and several variants of the method have been later introduced (Ba et al., 2016; Salimans and Kingma, 2016).

Ba et al. (2016) modify batch normalization by calculating the mean and variance estimates over the input features of a single example, instead of over the feature axis of a batch of examples and refer to this method as *layer normalization*. This removes the dependencies between examples within a batch and eliminates the sensitivity to a small batch size and is the most often used normalization variant in Transformer architectures.

2.3 Transfer learning

Whereas the neural architectures and optimization algorithms have had a strong impact on the progression of state-of-the-art NLP methods, one key concept can be seen as the revolutionizing factor in NLP during the past decade: transfer learning (Torrey and Shavlik, 2010). In essence, transfer learning frameworks focus on learning knowledge relevant for a given task and subsequently using this knowledge to solve other, often similar, tasks. Conceptually, it can thus be considered alike to (sequential) multitask learning, but often in transfer learning the ultimate goal is to solve the latter task, whereas the initial training task is only used as a support, e.g. to reduce the required training data amount for the actual task.

In language technology transfer learning was popularized by the pretraining of word embeddings. In these methods a dense representation for each word in the selected vocabulary is learned in a preliminary optimization task, which usually focuses on modelling the surrounding context of the words in a large corpus. As the context, and thus a large corpus of plain text, is the only required data to optimize these pretraining tasks, they are often called unsupervised, but from ML perspective they are very much supervised with target labels generated from the text corpus.

The first popular word embedding pretraining method, called word2vec, was introduced by Mikolov et al. (2013a,b). In the most common setup, originally coined as continuous skip-gram model, each unique word is represented with a one-hot encoded vector, which is then fed through two fully connected layers. The output layer size also corresponds to the size of the vocabulary. Each training example is produced by selecting a single word occurrence from a large text corpus and a neighbouring word from a small surrounding context. For the given input word, the task is to predict the word appearing in the context and the model can be optimized for this task using SGD based methods. This process is repeated for all input word - context word pairs in the corpus. As the output layer is shared between all input words, the only (input) word-specific learning can happen in the first fully connected layer. If two words are to often appear in similar contexts, i.e. the network has to predict similar neighbouring words for the input words, it is intuitive that the model will be learning similar weights for these words in the first fully connected layer [3]. Thus, words with similar semantic and syntactic roles tend to have similar word embeddings, i.e. the weights of the first fully connected layer. After the pretraining, this first fully connected layer can used as the first layer in other task-specific model architectures, leading to the semantic and syntactic word information being transferred to the new model.

Although word2vec became exceedingly popular, the concept of neural word embeddings and their behaviour as dense knowledge representations was introduced

[3] Word-specific weights here refer to the weights of the first fully connected layer attached to the non-zero input node.

at least a decade earlier. Outside the scope of neural networks, dense word representations for instance in the form of latent semantic analysis and random indexing have existed even longer (Deerwester et al., 1990; Kanerva et al., 2000). One of the earlier mentions of neural word embeddings, or distributed word representations as they were called, is in Bengio et al. (2003). Their study focused on neural language models, i.e. on the task of predicting the next word when the previous words are known. The model architecture shares many similarities with word2vec, i.e. it consists of a word embedding layer and a fully connected output layer, but the input can be multiple previous words. Whereas Bengio et al. (2003) did not transfer the word embeddings to a new task, they already mentioned the possibility.

Although word2vec was not the first nor the most complex word embedding method, its main contribution was to show that a simple model, with several computational optimization tricks, leading to the possibility of using much larger datasets can demonstrate much stronger performance than more complex methods. Indeed, (Bengio et al., 2003) trained their language model with a cluster computer and a corpus of 14 million word occurrences for 3 weeks. Mikolov et al. (2013a) on the other hand were able to train word2vec on a single CPU using a corpus of 1.6 billion tokens in a day. Such a vast computational performance improvement made it possible for the majority of researches to train their own word2vec models with domain and language specific corpora, ultimately leading to the popularization of pretrained word embeddings in NLP.

2.3.1 Deep transfer learning

Notice that with word2vec-like methods a dense vector representation (word embedding) is produced for every word, but the projection from a one-hot encoding to this word embedding is produced by a single fully connected layer. When this knowledge is transferred to a downstream task, only the first layer of the task-specific neural model is initialized with the word representations, whereas all other layers have to be learned from random initialization with the task-specific training data. Thus, in this thesis we refer to these methods as *shallow* transfer learning. As the rest of the model architecture tends to be more complex and deeper than the first word embedding layer, much of the learning relies solely on the task-specific data.

Another issue with word2vec-like word embeddings is that a single embedding is formed for a word, ignoring multiple word senses and contextual changes in the meaning. As an example, the word *Apple* will have a single dense representation, whether it refers to a fruit or a technology company.

The subsequent pretraining methods have thus expanded the concept of word embeddings in two ways: the pretrained models are deeper and form word representations for a full sequence, e.g. a sentence, thus conditioning the word embeddings on the context. These methods are often called *contextualized word embeddings* and

in this thesis referred to as *deep* transfer learning.

ELMo (Peters et al., 2018) was one of the first deeper transfer learning approaches for word representations. ELmo relies on stacked bidirectional LSTM layers, which are pretrained on bidirectional language modelling task, where the forward LSTM stack is trained as a standard language model and the backward LSTM stack on a reverse token sequence. For each token the model thus forms a representation based on its previous and subsequent tokens. The final encoding for each word is produced by taking a weighted sum of all stacked LSTM states. It is good to note that Peters et al. (2018) utilize ELMo in downstream tasks, by concatenating the produced word encodings with context-independent word embeddings and feed the concatenated representations to a task-specific RNN network. Thus, the models still contain complex components, such as RNN layers, which have to be trained solely on the task-specific training data. In practice the pretrained LSTM stacks are also fairly shallow, e.g. containing only two BiLSTM layers in the original paper.

Later models have taken the transfer learning even further: for instance BERT (Devlin et al., 2019) models are based on 12 or 24 stacked Transformer blocks pretrained on a masked language modelling (MLM) task, where arbitrary tokens in the input sequence are masked and the model is asked to predict the missing words. In addition, BERT uses a secondary pretraining objective, which focuses on an entailment-like task of predicting whether a given pair of sentences naturally follow each other in a larger text. However, the importance of this pretraining task has been disputed in subsequent studies (Liu et al., 2019). BERT demonstrated state-of-the-art results with a single fully connected layer added for task-specific finetuning. Thus, models initialized with BERT tend to have very limited number of weights that have to be learned from randomly initialized states and by relying only on the finetuning data, whereas most of the network is already trained with the MLM task before finetuning the model for the given task.

Although BERT is the only variant of Transformer-based models studied in this thesis, several variants of the model exist. For instance ELECTRA (Clark et al., 2020) simplifies the pretraining objective into a token-wise binary classification task, where the intention of the model is to predict, whether the given token has been replaced with another one. A generative masked language model is initially used to produce the input sequences with certain tokens replaced. In this sense the model resembles generative adversarial networks (Goodfellow et al., 2014), but the generator does not rely on an adversarial loss. Clark et al. (2020) report that this approach leads to comparable downstream task results as BERT's MLM pretraining, but with much smaller computational resources.

It is good to point out that although Transformer-based transfer learning methods have become the most popular approach and demonstrate overall best performance at the moment, in some tasks RNN models are still able to attain similar results. For instance FLAIR (Akbik et al., 2018, 2019) has demonstrated state-of-the-art re-

sults for named entity recognition with an LSTM-based architecture pretrained with character-level language modelling.

2.3.2 Relation to linguistics and philosophy of language

Whereas here the transfer learning methods have been described mostly from technological perspective and the advancements in NLP have been attributed to the machine learning researches developing such methods, it should be reminded that these approaches lean on prior linguistic theories proposed decades ago.

The underlying concept of all aforementioned pretraining techniques is that the semantic characteristics of a word can be derived from its context, i.e. the meaning of a word is defined by the way it is used in language. This fundamental thought makes it possible for these methods to learn powerful knowledge representations for words, phrases and sentences without any other supervision than the context of the targeted word.

This *distributional hypothesis* has been studied for instance by Harris in his work on distributional structure of language (Harris, 1954), where he argued that the distributional differences in contexts in which certain words occur also reflect the difference in meanings of these words, although he also argued that there is no parallel structure of meanings, which would strongly follow the structure of language. The same idea has also been crystallized in the quotes ”You shall know a word by the company it keeps!” by JR Firth and ”the meaning of words lies in their use” by Ludwig Wittgenstein (Firth, 1957). For Firth the concept of context, however, is much larger, incorporating the surroundings and events in which the language is used, and he explicitly mentions that context should not be confused with collocation. In language technology, this simplification is nevertheless done without exception and the connection of transfer learning and linguistic theories is not further discussed in this thesis. It is worth noting though that the size of the context has been growing in language models: with word2vec it has been common to use a window of roughly 10 tokens around the target token, whereas most Transformer models utilize context of hundreds of tokens. Lately multimodal models have also incorporated sensory data for grounding the language models, e.g. PaLM-E (Driess et al., 2023) utilizes visual and robot sensory data in addition to text.

3 Overview of the research

This chapter describes the goals, methods and findings of the studies included in this thesis. The chapter is divided into three distinct subsections based on the characteristics of the investigated research problems. The first section focuses on text classification (Aggarwal and Zhai, 2012), the task of labelling a given text snippet with the correct subset of predefined categories. The second section targets named entity recognition, the task of identifying relevant phrases within text. Named entity recognition (Mikheev et al., 1999) is often solved as a word-level sequence classification task, i.e. for every word in a text snippet a label has to be predicted. In addition the reduction of the named entity recognition task into a text classification problem is discussed, connecting these two distinct problem definitions. The third section addresses the task of predicting the molecular function of a given protein (Radivojac et al., 2013). Whereas this task seems unrelated to the natural language processing problems and techniques introduced in this thesis, these tasks in fact share a fundamental similarity: modelling written language as a sequence of characters or words is analogous to modelling a protein as a sequence of amino acids. Thus an important research question in the third section is whether latent neural representations used in language technology are also suitable for solving tasks in the biological domain.

3.1 Text classification

This section describes experiments on clinical text classification, focusing on two distinct applications. The first one focuses on thematic (subject heading) classification of nursing notes on sentence level, whereas the second one targets social media texts in an attempt to identify medication intake events. There are three distinct features in these tasks: (1) the former study is conducted on Finnish and the latter on English texts, (2) the nature of texts in clinical and social media domains differ drastically and (3) the first task is strongly multiclassed with hundreds of possible labels, which are not necessarily mutually exclusive, whereas the latter task is restricted to three exclusive classes.

Despite the differences, the tasks also share several similarities. Both can be considered as short text classification tasks and can be addressed with techniques developed particularly for such problem settings. In addition abbreviated and grammatically incorrect expressions, if such normative stance is permitted, are common

for both of the studied domains (Haverinen et al., 2009; Liu et al., 2012).

We describe both of these tasks in detail, explaining their importance, the characteristics of the data and the conclusions in their dedicated subsections. The study on Finnish nursing note classification has been published in papers I and II, and the medication intake detection in paper III.

3.1.1 Related work in text classification

Text classification is a well studied task with a vast range of publicly available benchmark datasets (Chakraborty et al., 2016; de Gibert et al., 2018; Pang and Lee, 2005). The GLUE and SuperGLUE natural language understanding benchmarks (Wang et al., 2018, 2019), which focus on the general effectiveness of methods over a wide range of tasks, have become the boxing ring for the Bertian methods. Both of these benchmarks also contain multiple tasks which are either inherently text classification or can be formalized as such. At the time of writing, these benchmarks are dominated by such Transformer (Vaswani et al., 2017) variants as ERNIE (Sun et al., 2021), T5 (Raffel et al., 2020) and DeBERTa (He et al., 2021). ERNIE focuses on pretraining where information from a knowledge graph is incorporated with the standard masked language model task while T5 converts all tasks, including text classification, into text generation. DeBERTa on the other hand changes the model architecture by disentangling positional information from the contextualized word embeddings. None of these studies pay closer attention to the fine-tuning on a given task, but instead highlight the importance of transfer learning. Notice though that all of these models have been released years after the studies conducted in this thesis.

Earlier methods in text classification have relied on e.g. support vector machines and TFIDF features (Joachims et al., 1999). After the discovery of efficient word embedding learning, e.g. word2vec (Mikolov et al., 2013a), the classification methods started to incorporate these semantic representations as features. However, support vector machines still remained as a state-of-the-art method and the word embeddings were often simply summed or averaged to form a feature representation for a text sequence (Lilleberg et al., 2015). As this naive approach led to only minor improvements in text classification performance, subsequent studies have focused on sequential modelling approaches, often in the form of convolutional (Kim, 2014; Zhang et al., 2015) or recurrent neural networks (Lee and Dernoncourt, 2016; Liu et al., 2016).

In the biomedical and clinical domain text classification experiments are also common (Pestian et al., 2007; Kavuluru et al., 2015; Mujtaba et al., 2019; Hughes et al., 2017). However, for Finnish nursing notes the prior studies are scarse and mostly conducted by Suominen and Salakoski (2010), Suominen et al. (2006) and Hiissa et al. (2007). To my knowledge no prior text classification studies have been conducted on Finnish nursing notes in the neural NLP era. The situation is simi-

lar for the vast majority of languages as a recent review suggests that 91% of deep learning based clinical natural language processing studies focus on either English or Chinese datasets (Wu et al., 2020a). Whereas drawing conclusions of general model performance based on a single evaluation task is shown to be challenging even in monolingual settings (Zhang et al., 2015), doing so is even more challenging when estimating the performance of a method for a new target language. Thus it is critical to not only measure the performance of new approaches on large and well-resourced languages such as English, but to also measure whether these approaches deliver similar performance on other languages. Moreover, using multilingual resources can also improve model performance on high-resource languages (Duque et al., 2016), warranting the creation of datasets in various languages even for the purpose of improving model performance on English.

Our study on medication intake mentions is conducted in the context of a shared task (Sarker et al., 2018) and the related work on this specific task is discussed in Section 3.1.3.

3.1.2 Thematic classification of Finnish nursing notes

Care documentation is a mandatory process for safe care continuity. Such documentation usually contains information such as patient's current health status, previously administered care and future treatment plans. Often a specific structuring of the textual documentation is required to enable an easy access to the essential information at any given time. Due to the complexity and exhaustiveness of the documentation, prior studies have shown that up to 35% of nurses' working time is spent on documentation (Yee et al., 2012).

In this study the data is collected from a Finnish hospital district where the documentation is manually divided into thematically coherent paragraphs. The subject headings of these paragraphs should follow the Finnish Care Classification (FinCC) ontology, but this is not enforced by the information system. Thus, nurses are presumed to memorize hundreds of ontology concepts and to be able to select the most suitable concept for each paragraph. The headings are also written in free text, allowing the creation of new headings not present in the ontology, but also leading to a multitude of misspelled variants. The data also contains headings from past care documentation ontologies before FinCC was adapted. Moreover, this documentation practice breaks the natural narrative and chronological order of events as each sentence must be assigned to the corresponding thematic paragraph, whereas other concurrent observations are listed under other headings, obscuring their possible relations.

Our aim is to automate the heading assignment on sentence level. That is, we wish to allow care documentation in a free narrative manner and subsequently restructure the document into thematic sections if necessary. We believe this to have

several benefits for both care documentation and information retrieval. Firstly, the approach has the potential to speed up the documentation process as nurses are no longer required to manually structure the documentation, freeing up nurses' time for providing actual care. This benefit should be even more pronounced if in the future the documentation can be produced by dictation using speech recognition models. Secondly, such system can standardize the usage of controlled ontologies, improving the coherence of documentation across different hospital units.

From machine learning perspective this task setting has a few interesting properties. Gold standard labels are transferred from paragraphs to sentences with the underlying assumption that sentences are correctly assigned to the thematic paragraphs. This is a similar assumption as is made in distantly supervised approaches and inevitably leads to noisy data. Another source of noise is that the headings are assigned by nurses over a long span of time under constantly changing care practices, usually under strict time constraints and without proper familiarization with the used ontology. Lastly, each sentence only contains a single corresponding heading, but these headings are not mutually exclusive. Thus the task resembles multilabel classification, but for training and automated evaluation purposes we only have multiclass data.

Data and Methods

The used nursing note data originates from the EHR system of a Finnish university hospital, covering half a million nursing notes. In this experiment, only headings with at least 100 corresponding sentences were included, resulting in 676 unique headings and 5.5 million sentences, divided into training, development and test sets. Figure 4 shows an example of a nursing note and Table 1 lists some of the most frequent and infrequent headings.

We benchmark several text classification methods on this data, including recurrent, convolutional and feed-forward neural networks, as well as more traditional linear classifiers with bag-of-words feature representations. For the RNN (LSTM) and CNN approaches pretrained word embeddings are utilized. Detailed descriptions of all tested methods are in Paper I.

Automated evaluation results

We start the evaluation by automatically comparing the predictions against the headings from the original nursing notes in the test set, which contains over million sentences in total. As the primary measurement we use accuracy. We are also motivated by a use case where the system is not used in a fully automated fashion, but instead assists the documentation process by suggesting suitable headings while the nurses are documenting the care. For such a use case finding the exact matching heading is

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

NUTRITION
Eaten breakfast.
Eaten lunch.

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

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

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

EDUCATION OF RELATIVES
Wound treatment instructions and pain prescriptions given.

Figure 4. A fictional depiction of a nursing note structured under subject headings following the writing style found in the studied Finnish patient records. Source: (Moen et al., 2018)

not necessary, but it is sufficient that that model is able to rank the correct heading highly. Thus, we also measure the classifiers' recall and mean reciprocal rank when the 10 highest confidence headings are selected. These metrics are often seen in recommender system and information retrieval evaluations (Herlocker et al., 2004; Shi et al., 2012; Voorhees et al., 1999).

According to this evaluation (see Table 2) the LSTM based models are the best choice for the task, with a minimal improvement provided by adding bidirectionality to the recurrent layer, resulting in accuracy of 54.4%. The evaluated convolutional model provides also competitive performance with accuracy of 53.5%, whereas a linear SVM baseline results in accuracy of 51.5%. Notice that due to the large num-

Subject heading	Count	Percent
Wellness and Ability to Function	222,984	6.737
Physiological Measurements	198,919	6.010
Nutrition	135,984	4.109
Urinary Tracts	128,486	3.882
Activity	123,294	3.725
Medication	117,502	3.550
Urinary Incontinence	101	0.003
Change in the Kidney and Urinary Tract Activity	101	0.003
Other	101	0.003
Neuropathic Pain	100	0.003
Coping with Activities of Daily Living	100	0.003
Loss of Appetite	100	0.003

Table 1. The most and least common headings included in the dataset. The heading names are translated from Finnish. Source: Paper I

Method	Accuracy	R@10	MRR
BidirLSTM	0.5435	0.8954	0.6621
LSTM	0.5429	0.8932	0.6610
CNN	0.5348	0.8856	0.6526
fastText	0.5224	0.8801	0.6428
BoWLinearSVC	0.5149	0.8486	0.6286
RandomForest	0.4896	0.7690	0.5868
Word&HeadingEmbeddings	0.1629	0.5111	0.2633
MostCommon	0.1038	0.3776	0.1679
Random	0.0015	0.0150	0.0044

Table 2. Automated evaluation results of the selected methods. All performance differences between methods were found to be statistically significant. Source: Paper I

ber of headings, random baseline has only 0.2% accuracy and selecting the most common heading results in accuracy of 10.4%.

Assessment of predictions and data quality

The automated evaluation suggests that all the evaluated models have relatively poor performance in comparison to various other text classification tasks (Zhang et al., 2015). We speculate that this does not actually reflect the ability of the models, but instead the experimental setting and data have two major challenges. Firstly, we train and evaluate the model in a multiclass setting, where subject headings are considered mutually exclusive. Many of the used subject headings are in fact hypernyms and hyponyms of each other resulting in various headings being suitable for a given sentence. However, the EHR system does not enforce direct linking to the FinCC ontology and we are unable to trace the hierarchical relations of the subject headings. Thus, this possibility is not considered in the automated evaluation.

Secondly, the gold standard headings are fetched directly from the university hospital EHR system without any additional curation or quality control. Thus, the data may contain considerable amount of noise due to erroneous heading assignment or slightly varying documentation conventions across different hospital units.

To assess these hypotheses we conduct identical manual evaluation of both the original and predicted headings (from the best performing bidirectional LSTM model). For this evaluation a subset of 200 random sentences were selected and the correctness of the assigned headings were evaluated by 3 domain experts.

The evaluation of the original headings reveals that the domain experts considered only 74% of the cases to be assigned under correct headings leading to a considerable amount of noise in the training data. This inevitably leads to artificially deteriorated performance scores in the automated evaluation.

The most surprising result of this study stems from the manual evaluation of the predicted headings as the evaluation of the predicted headings on the same subset of sentences resulted in 82% of the headings to be correct. That is, even though the model performance was measured to be only on the level of 54% in the automated evaluation, the domain experts preferred these predictions over the original headings assigned by nurses. In addition, 13% of the headings were considered possible correct, but the curators were not fully confident in their judgements without seeing the full context of the sentences. Thus, optimistically speaking, it is possible for the prediction accuracy to be as high as 95%.

Model performance on free narratives

As the nurses are mandated to write the care documentation under a preassigned subject heading, the sentences tend to contain information related to a specific concept,

breaking the narrative and relations to other events and concepts. As such, we do not expect the previously discussed experiment to realistically reflect the systems' performance if the nurses were allowed to dictate or write the care documentation without any restrictions regarding the structure. Thus, in paper II we conduct experiments on how well the bidirectional model performs on free narrative texts. As no such care documentation exists in the university hospital EHR system, for this experiment patients with artificial backgrounds were created and a group of nurses were requested to write fictional care documentation for one day of care for each patient in free narrative style. In total 20 such daily nursing notes were constructed, resulting in 457 sentences.

The bidirectional LSTM model was subsequently applied to the written sentences for subject heading assignment. These predicted subject headings were then manually evaluated by domain experts in a setting identical to the evaluation in Section 3.1.2. According to this experiment the accuracy of the model was 68%, clearly inferior to the 82% measured on the semi-structured data. However, the portion of possibly correct headings was found to be larger (20%), suggesting that the model is still very capable of suggesting relevant headings.

This experiment demonstrates the difficulty of transferring a model trained on semi-structured texts to free narratives, even within the same domain. We speculate that whereas pre-selecting the subject heading strictly forces the nurses to write about a single topic in each sentence, this assumption is often violated in free narratives, where language use is more compressed and casual, leading to sentences discussing multiple topics and thus confusing the classifier trained on a multiclass objective. This issue was also mentioned by the domain experts conducting the evaluation and is probably reflected in the larger number of possibly correct headings, as the predicted headings may be correct for a part of a sentence, but do not describe the whole content.

Coherence of heading predictions

For the purpose of automatically restructuring a free narrative under suitable subject headings, it is crucial that the used text classification model is relatively consistent and not too specific to prevent each sentence from being grouped into a separate category. Too specific categories will lead each sentence falling into its own category and if the hierarchy of the categories is not known, they cannot be aggregated later on. In Paper II we evaluate the predictions of the bidirectional LSTM model from two additional aspects: whether the paragraphs formed by grouping the sentences based on the predicted subject headings are sensible and whether the predicted subject heading is suitable for the whole paragraph, i.e. we move from a sentence classification task towards the actual restructuring of a free narrative nursing note, which is a primary goal of this study.

We also introduce a post-processing step to further merge paragraphs with similar subject headings. In this step every formed paragraph (*source*) is compared against the other paragraphs and potentially merged with a *target* paragraph if they are found to be similar enough. This merging is based on the similarity of the latent subject heading representations (see Section 3.1.2) of the compared paragraphs, the classifier confidence of the sentences in the *source* paragraph belonging to the *target* paragraph and the number of sentences in both of these paragraphs.

The constructed paragraphs were manually evaluated by domain experts. Based on this evaluation grouping the sentence simply based on the predicted headings results in paragraphs out of which 66.67% were coherent and also had a descriptive subject heading. With the additional post-processing step this portion rose slightly to 68.85%, suggesting that merging the original paragraphs does not decrease the quality of the restructured nursing notes. However, the number of paragraphs is reduced by 23%, as the too specific headings are grouped together. Notice that in some cases the paragraphs may seem coherent, i.e. the grouped sentences discuss the same topic, but the selected subject heading may not be descriptive of the paragraph. The more detailed evaluation of these cases is discussed in Paper II.

Latent heading representations as indicators of actualized ontology use

We hypothesize that the FinCC ontology contains too detailed headings for practical purposes, hindering the nurses' ability to correctly select the most suitable heading while documenting different aspects of the administered care and possibly also slowing the documentation process.

The used neural network models can be considered to encode each input sentence into a single vector and the final fully connected layer relates the encoding to all possible headings. Such division of the roles of the components of the neural model, albeit artificial, is for instance used also in the word2vec model.

As the hidden layer parameters are shared among all target classes (headings), on the output layer, each heading has a set of dedicated weights. Thus, it is intuitive to expect that these output layer parameters represent the semantics of each heading and can be used to measure the similarities of the headings. As the model is trained on labels assigned by nurses during their work, these semantic representations in particular represent the meaning of the headings as perceived by the nurses, not the normative meaning assigned by the creators of the ontology.

Thus we further analyze how the model relates different headings to each other and whether this reflects the FinCC ontology. We start by visually inspecting the hierarchical clustering of the heading representations. A subtree of the formed clustering dendrogram is shown in Figure 5. This subtree contains two separate high level clusters, the first one focusing on concepts related to breathing and the second

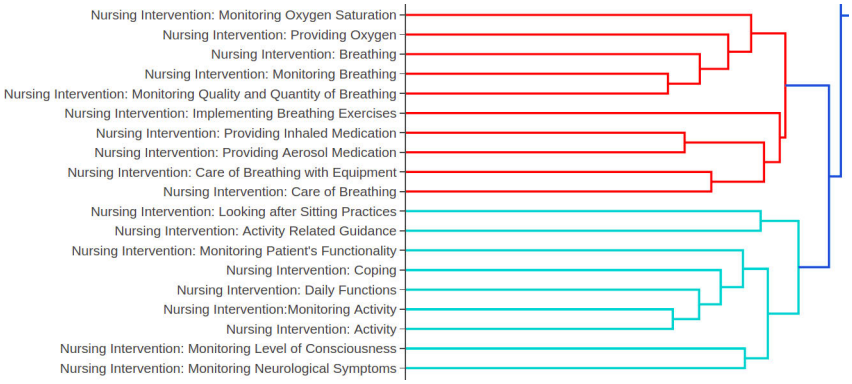

Figure 5. A subtree of the heading dendrogram formed with hierarchical clustering of the heading representations derived from the classifier. The headings have been translated from Finnish. source: Paper II

on general activity of the patient. The overall quality of the clustering is surprisingly good. For instance, concepts *Monitoring Breathing* and *Monitoring Quality and Quantity of Breathing* as well as *Providing Inhaled Medication* and *Providing Aerosol Medication* form the nearest pairs in the clustering. The visual inspection strongly encourages the utilization of these representations.

Whereas the heading representations extracted from the classifier can be compared by measuring the cosine or euclidean distance of the vectors, we measure the similarity of headings in the original FinCC ontology by calculating the shortest path between the given concept pair in the hierarchy. This simple method has been used for other biomedical ontologies as well (McInnes et al., 2009). Thus, we can measure similarity for each possible concept pair found in the FinCC ontology or in the classifier output layer. Each pair is subsequently ranked based on these similarity scores, resulting in two independent rankings.

The first observation from these rankings is that Spearman's rho between them is only 0.12, suggesting that these similarity measures do not show strong correlation. Whereas this could be caused by the neural model learning poor semantic representations for the headings, its classification performance as well as the manual inspection of the clustered headings suggest otherwise. Thus, we focus on concept pairs where these similarities strongly disagree. It turns out that all top 1000 pairs with the largest difference in the ranking, in the direction in which the classifier considers them similar, but the shortest path does not, originate from pairs where one concept belongs to the *Nursing Diagnosis* and the other to the *Nursing Intervention* top level categories, i.e. the root of the ontology tree is the only common ancestor.

As an example, the most conflicting cases include pairs such as *Nursing Diagnosis: Swelling – Nursing Intervention: Monitoring Swelling* and *Nursing Diagnosis: Changes in Oral Mucosa – Nursing Intervention: Basic Care of Oral and Other*

Mucosa. As such, these concepts essentially enforce the nurses to separate the administered care from the related observations, or the classifier is not able to clearly distinguish such small differences. However, to support the former hypothesis, we have also looked into the sentences written under these headings. For instance the *Nursing Intervention: Monitoring Swelling* heading contains sentences such as *Shins somewhat swollen* and *Shins still swollen, feet not as much*. These sentences are clear observations of the patient's status, i.e. should be documented under the *Nursing Diagnosis: Swelling* category, not under the *Nursing Intervention*. Moreover, the data includes word-for-word identical sentences, such as *Legs swollen* under both of these headings, suggesting that nurses are unable or unwilling to differentiate between these concepts.

As a result, we argue that the neural models provide valuable information about the perceived semantics of the underlying concepts and the latent knowledge representations can be utilized in assessing the suitability of the used ontology, for instance by pinpointing the problematic concepts. These findings could further be used in developing the future versions of the FinCC ontology or in the documentation training provided for the nurses. However, note that this is not an in-depth analysis, but merely a preliminary proof-of-concept demonstrating that neural models are not applicable only for the task they are trained for, but provide meaningful insights into language and ontology use. Further research is needed to establish quantitative measurements for evaluating the model in such applications.

3.1.3 Medication intake in social media

The second text classification task studied in this thesis focuses on utilization of social media, namely Twitter, content in pharmacovigilance. Such a vast source of information can have a vital role in detecting rare adverse drug effects, not necessarily detectable in small scale clinical trials, but the noisiness of the data poses severe challenges for real applications.

Instead of directly assessing the detection of adverse drug events, in this study we focus on a related task of detecting mentions of actualized medication use. In this task the goal is to automatically recognize whether a user posting to social media is describing their own medication use, which has already happened, in contrast to mentioning drug names in general or describing intended future use. From the pharmacovigilance perspective this classification step is crucial in reducing the amount of noise in the data, as a large portion of the drug mentions in social media are not related to the personal medication intake experiences of the content creator.

These experiments were conducted in the framework established in the Social Media Mining for Health Applications shared task (Sarker et al., 2018). The main advantage of such a shared task is a unified evaluation platform and reduced efforts for establishing baselines as each participant is able to focus in their own approach.

Class	Examples
Intake	Advil has been my bestfriend lately OMG.
	@_JasminePatrice oh i did cry , everyday lol but i had to
	take Motrin cause my hydro's had me too drowsy w/ the baby here ..
Possible intake	Tylenol and Xanax to numb me up fuck the bs lol this pain ain't no joke
	When Jocelyn brings me Advil to work
Non-intake	Prozac nation is so good why haven't I watched this before
	Xanax flow is rl oc af

Table 3. Example tweets from each class considered in the medication intake classification task. Some of the messages demonstrate the non-standard language use and vocabulary common in social media.

Thus, in this study we have conducted a limited number of more focused classifier training experiments, instead of comparing a vast amount of alternative approaches as was essential for the subject heading classification task. The presented methods are instead compared against approaches developed by other participating teams.

Experimental setting and data

The data used in these experiments is provided by the organizers of the shared task and originally consisted of over 10,000 manually annotated English tweets, all of which mention a drug name. Each tweet is assigned to one of three classes: *intake*, *possible intake* and *non-intake*. The first class contains tweets which clearly express personal intake, whereas the *possible intake* is more ambiguous, yet still suggests an intake event by the tweet writer. The last class contains a miscellaneous group of tweets, which mention a drug, but do not imply medication intake by the writer. These can be for instance medication suggestions for other people or general discussion on recreational use of drugs. Example sentences from each class are shown in Table 3. Approximately half of the data belongs to the *non-intake* class, one third to the *possible intake* class and one fifth to the *intake* class.

The data was released as a set of tweet IDs and their corresponding annotated classes. Thus, it was mandatory for the participants to download the textual tweets and their metadata themselves. This resulted in somewhat flawed experimental setup where the data was partially lost after the initial release. In our case, 7% of the training data was unattainable, making model comparison across different research groups difficult.

The task is considered as a standard multiclass classification experiment and the official evaluation uses micro-averaged F-score of *intake* and *possible-intake* classes as the primary metric. In our experiments, we follow the same evaluation approach.

Hyper-parameter	Optimal value	Tested Values
Character embedding dimensionality	25	[25,50,75,100]
Word embedding dimensionality	400	pre-trained
Character CNN, number of filters per window size	50	[50,100,150,200]
Character CNN, window sizes	[2,3,4,5]	[2,3,4,5]
Word CNN, number of filters per window size	200	[100,200,300]
Word CNN, window sizes	[2,4]	any subset of [2,3,4,5]
Dimensionality of first dense layer	400	[100,200,300,400,500]
Dropout rate	0.2	[0,0.2,0.5]
Activation functions	tanh	[ReLU, tanh, sigmoid]

Table 4. The optimal and tested hyperparameter values of the CNN-based system.

Methods

In this experiment we focus on a convolutional neural network model, although other standard neural architectures were also briefly tested. Unlike in the clinical text classification project, in this experiment we also use character-level modelling. Each input tweet is represented as a sequence of words and as a sequence of characters. All elements in both of these sequences are represented as learnable latent vectors (embeddings). The word embeddings are initialized with word2vec vectors, pre-trained on a large set of drug related tweets. Character embeddings on the other hand are randomly initialized and only trained on the actual task-specific data. Independent sets of 1D convolutions are applied on both of these sequences. The output of each convolutional kernel is further max-pooled across timesteps for each kernel independently leading to a vector-shaped encoding of the input sequence. The vectors from word- and character-level convolutions are subsequently concatenated and fed through two fully connected layers to form the final prediction. We tune the model hyperparameters with a grid search and all tested hyperparameters are listed in Table 4.

As a baseline method we use a linear SVM classifier with TFIDF weighted BoW representations. As this approach resulted in strong performance (see Section 3.1.3), we decided to incorporate similar features to the neural model. The resulting model receives as an additional input the BoW representation of the document, but the original sparse representation is compressed to a 4000 dimensional dense vector using truncated singular-value decomposition (SVD) (Halko et al., 2011), similarly to latent semantic analysis (Deerwester et al., 1990; Dumais, 2004). The full model architecture is visualized in Figure 6.

Since we found the training to be relatively unstable with the selected hyperparameters, the final model is constructed as an ensemble of a set of CNN classifiers. All of these models are identical, following the description of this section, but use different random initialization values for parameters. For prediction, the confidence values from these models are summed and subsequently the argmax is selected as the

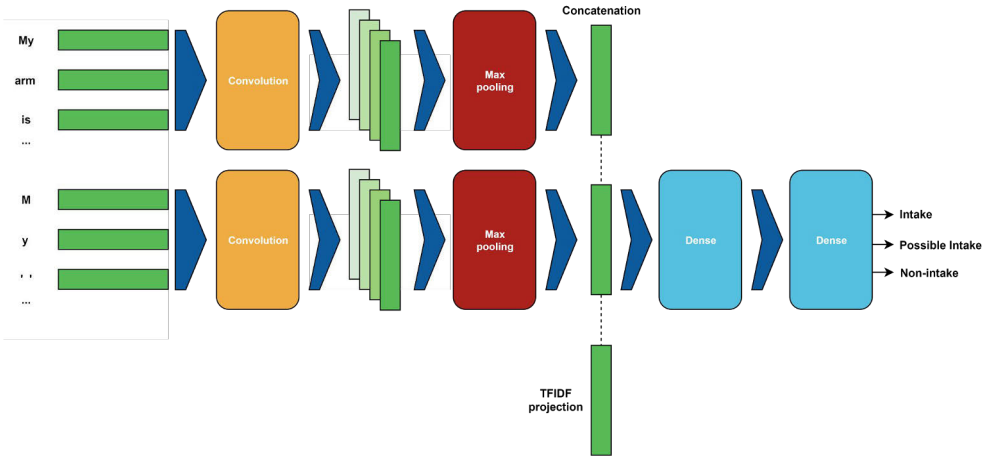

Figure 6. The neural model used in the medication intake classification task.

predicted class. In total 15 of such networks were trained, but only the subset maximizing the model performance on the development set is used as the final model, leading to an ensemble model of 6 networks.

Results and discussion

The overall results of our approach are listed in Table 5. The primary performance metric is the micro-averaged F-score of *intake* and *possible intake* classes, following the official evaluation. Notice though, that due to the inaccessible data points, the development set results are not directly comparable to other studies. However, the test set results are presented as reported by the shared task organizers.

Our CNN approach surpasses the SVM baseline by 2pp in F-score on the test set, showing relatively strong performance with an overall F-score of 66.3%. Although the *intake* class has more crisp definition in the annotation guideline, the performance of the classifiers is actually weaker for this class than for the *possible intake* class. This is most likely caused by the data distribution, *intake* class being slightly less common.

Out of all the shared task participants, the best performing system was created by the InfyNLP team (Friedrichs et al., 2018) and thus we have added their test set performance for comparison. InfyNLP system outperforms our CNN model by 3pp in F-score, showing remarkably strong performance. Interestingly, their system shows several similarities with our model. Their basic building block is a shallow CNN classifier, identical to ours: a sequence of words is represented with embed-

		Development set			Test set		
		Precision	Recall	F-score	Precision	Recall	F-score
SVM	Intake	70.5	64.5	67.4			
	Possible Intake	73.3	68.6	70.9			
	Overall	72.3	67.0	69.6	69.2	60.1	64.3
CNN	Intake	70.9	71.3	71.1			
	Possible Intake	76.3	71.1	73.6			
	Overall	74.2	71.2	72.7	70.1	63.0	66.3
InfyNLP	Overall				72.5	66.4	69.3

Table 5. Overall performance of our SVM and CNN-based systems. The development set results are measured with our own evaluation whereas the test set scores are as reported by the organizers. The class specific performance was not evaluated by the organizers and has been thus left out from the table for test set. For comparison we have added the results of the best performing team: InfyNLP. Source: Paper III

dings and convolutional kernels are applied on top. Following the CNN layer, 2 fully connected layers are utilized. However, they do not use character or BoW based components like our model and adding these to their system might increase their performance even further. The second similarity comes from the fact that also their approach relies on an ensemble. However, their ensemble approach uses two levels of aggregation, which they call a stacked ensemble. In this system, they use a 5-fold cross-validation on the combined training and development sets and average the prediction confidences from these models to form the first ensemble of 5 classifiers. This process is then repeated 99 times and the top 20 ensembles are selected in the second aggregation level to produce the final model, totalling in 100 independent CNN models. Note that this should not be confused with the general stacking ensemble method (Wolpert, 1992), where the predictions from a group of models are passed as an input to another model.

The main advantage of the InfyNLP approach over ours, is most likely the use of cross-validation (CV), which allows InfyNLP to utilize the small dataset more efficiently. Although CNN models are computationally very efficient especially on GPU and TPU systems, it can be argued that running 100 such models to form a single prediction is not practical in many cases considering the gained benefits.

Whereas most of the top performing systems suggested for this task rely on neural approaches with pretrained word representations, it is good to note that more traditional approaches, such as our linear SVM baseline are still very competitive methods. Indeed, the approach suggested by team NRC-Canada, which relies on a linear SVM classifier and domain-tailored set of features outperforms our CNN model with an F-score of 67.3%.

As we only had access to a partial training data, we estimate how much the performance of the systems can be improved with additional data. In this experiment we train the CNN model with an incrementally grown subset of the training data,

starting from 5000 examples, and measure the performance gains. With each subset size, we train 5 independent models and report the mean performance. The relation of training data size and model performance is plotted in Figure 7. The performance gain seems fairly linear with no plateauing in this region, suggesting that the partial loss of the training data has an impact on our performance. By linearly extrapolating with the regression line fitted on the performance measurement, we expect to gain 0.7pp increase in F-score with the full training data. Based on this experiment it is also safe to assume that the CV approach used in the InfyNLP system also provides similar gains.

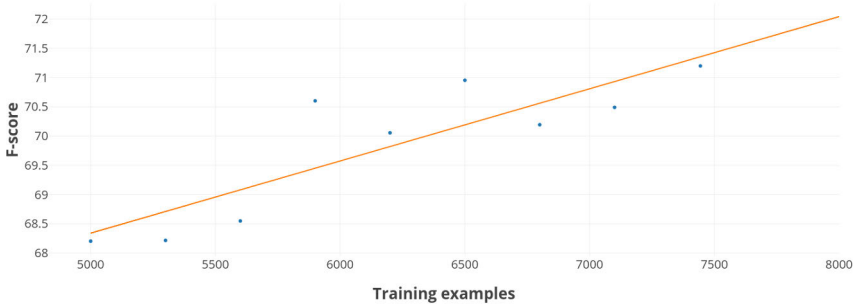

Figure 7. Influence of the number of training examples to the performance of the CNN system as evaluated on the development set. Source: Paper III

Although the organizers of the shared task provide extensive annotation guidelines for the dataset, some of the annotation criteria appear very challenging to interpret, most problematic case being the *possible intake* class. For instance ”If I take any more acetaminophen for my back my liver may explode....” is considered as a *possible intake*, whereas ”Ugh. Tylenol as a pregnancy pain reliever is such a joke.” should be annotated as *intake*. The reasoning for the latter annotation is given as: ”... given the real-time nature of social media and the absence of explicit temporal markers of the past or future. From this present orientation, we can infer that the medications were taken recently” (Klein et al., 2017b). On the other hand, for the former example, the guideline states: ”... only indicate that the authors take the medications frequently in general”. Detecting such small differences in tweets seems extremely difficult, at least from the perspective of a non-native speaker, although the organizers mention inter-annotator agreement of 0.88 in Cohen’s Kappa (Klein et al., 2017a), which can be considered remarkably high. To verify this, we manually annotated a subset of 100 tweets from the development set and compared the annotation against the gold standard. Our manual annotation resulted in an abysmal F-score of 59.3, notably lower than the performance of the developed methods, indicating that the task is indeed challenging for humans if the annotation guidelines are

not carefully followed with enough time to verify unclear cases.

The annotation guidelines also heavily focus on temporal markers and past-tense verbs as indicators of actual *intake* having happened. Indeed, past-tense verbs are twice as common in the *intake* tweets than in *possible intake* tweets. As for instance Penn Treebank POS schema represents past-tense verbs as their own word class, we also tested including part-of-speech (POS) tags as additional features to our model, treating them analogously to word and character sequences. However, we did not notice any performace gains from this experiment, probably due to the pretrained word embeddings already containing enough information of the similar syntactic behaviour of words within the same word class, rendering the distinct POS features unnecessary.

In addition to the final CNN based models, several other standard architectures, with no observed performance improvement, were tested, including RNNs. These LSTM networks were further incorporated with attentive capabilities. In this approach instead of only using the final LSTM output, the encoding from each time step were used and subsequently pooled. The pooling was conducted in an attentive manner, where another LSTM layer was used to produce the encoding the attentive pooling was conditioned on. Note that in this model a single attention head was used to pool all the time steps of the RNN encoder to form a vector shaped weighted mean encoding of the underlying word sequence. This differs from the commonly used and greatly successful self-attention mechanism, where a pooled representation is formed for each time step in the input sequence, conditioned on every time step separately. It is good to remind that this study was initiated before the publication of Vaswani et al. (2017), which introduced an effective way of fully self-attentive architectures in the form of Transformers, and as a hindsight their adaptation of self-attention seems far more intuitive than the model we experimented with. Yet, it took a notable amount of further research, before the Transformer model demonstrated state-of-the-art performance in text classification tasks, with deep pretraining of the network, in the form of BERT Devlin et al. (2019).

3.1.4 Summary and future work

In this section we have studied two distinct text classification tasks, both related to medical research, yet one focusing on clinical texts written by domain experts and the other on social media content. Some distinct features of these tasks are listed in Table 6. Most importantly, the medication intake classification task has a well defined framework and guidelines provided by the shared task organizers and represents a common text classification task, making the method development studies fairly straightforward. The clinical subject heading prediction on the other hand has several challenges due to the nature and origin of the data and their incompatibility with the ultimate goals of the project.

Dataset	**Subject headings**	**Medication intake**
Data source	Patient records	Twitter messages
Label count	676, hierarchical	3, mutually exclusive
Annotation source	Assigned by nurses during daily patient documentation	Two expert annotators with comprehensive annotation guidelines
Data amount	∼5.5 million sentences	∼10,000 sentences

Table 6. Comparison of the text classification datasets used in this study.

In both of the tasks, neural models demonstrate similar performance improvement over our linear SVM baselines: 3pp in accuracy for subject headings and 3pp in F-score for medication intake, suggesting that exploring such models is indeed beneficial in biomedical text classification. Nevertheless, the absolute performance in these tasks is quite different as the accuracy for subject heading prediction is only 54% with our best performing model, whereas F-scores of over 70% can be achieved for the simpler medication intake task. These measurements may not however be good indicators of the difficulty of the tasks, but instead the noisy nature of the validation data used in the clinical subject heading prediction task obscures the automatic evaluation.

To avoid making too rushed conclusions on the applicability of these methods, we have conducted manual evaluation of the subject headings, for both the original data as well as the predicted headings. Similarly we introduced a new set of annotations for a subset of the medication intake data. In both of these cases the neural models either reached or surpassed human performance, when given certain constraints for the manual annotation work. In the medication intake task, the annotator was not a native English speaker nor had he memorized all rules and exceptions stated in the guidelines, which resulted in our annotations agreeing with the gold standard much less than our trained models. More interestingly domain experts considered the predicted clinical subject headings more suitable for the given sentences than the original headings collected from the EHR system. The implication of this is that the insufficient time resources given to the nurses documenting their provided care, which usually makes it impossible to browse or reread the full FinCC ontology during documentation, leads to erroneous use of the terminology and to a poorer documentation quality. In the current state of language technology, it seems to be possible to fully automate this task with no compromises being made in the data quality, although we have also shown that if given the freedom to write completely unstructured narratives, the complexity of the produced documentation increases, resulting in a lower performance of the classifier, trained on the existing semi-structured data.

Notice that both of these studies were conducted before the introduction of *deep* transfer learning methods such as the Bertian models. However, my expectation is

that in the subject heading classification task where the training data is abundant and even the *shallow* transfer methods are able to reach human-level performance, similar leaps in performance as has been demonstrated in various benchmarks (Devlin et al., 2019), will not be achieved with newer methods. In the medication intake detection, on the other hand, the shared task organizers report very high inter-annotator agreement, suggesting that the task should be relatively easy for trained humans, leaving a large margin in performance between the suggested methods and humans, although the reported Cohen's Kappa is not directly comparable to the used evaluation metric. In this task considerable gains could be expected from more recent neural classification methods. Our experiments on varying training dataset sizes also suggest that higher performance on this task is achievable either through transfer learning or by collecting additional task-specific training data.

Whereas the success story of Transformers (Vaswani et al., 2017) has directed the focus of neural network research on model architectures relying on multi-head self-attention, these studies should be accepted with healthy criticism. Our classification studies suggest that architectural choices play relatively small role in the outcome: even though we have reported stronger results for RNN-based models compared to CNNs in the subject heading task, it is possible that this is caused by slightly more suitable hyperparameters. Indeed recent studies have similarly questioned the importance of the whole self-attention mechanism in the Transformer architecture (Tolstikhin et al., 2021; Melas-Kyriazi, 2021).

In the medication intake detection, we have demonstrated that neural models relying on *shallow* transfer learning are still able to benefit from more traditional bag-of-words features. The same idea is taken further in Topic Memory Networks (Zeng et al., 2018), where a neural auto-encoder is used to densify the BoW representation instead of SVD as in our approach. The created representation is subsequently incorporated with word embeddings in an attention-like manner, although the authors do not use this term. It is good to keep in mind that the self-attention architecture itself is a form of a weighted BoW representation as it does not inherently take into account the word order and this information has to be introduced via the positional embeddings, yet it does utilize dense word embeddings which are not orthogonal.

The major challenges in the subject heading classification arise from the mismatch of the multiclass training and evaluation data and the actual multilabel task that should be solved. A critical future work direction will be to solve how the classification models can be automatically evaluated with meaningful and interpretable metrics despite this discrepancy as our manual evaluations suggest that the studied models perform much better than can be expected based on the current automated evaluation approach.

As the text classification models show strong enough performance to be applicable in real life scenarios, the most crucial future objective is to take the leap from academic research into actual electronic health record systems. As of now, it is still

unclear how this should be done, and which aspects, such as the actual increase in productivity or saved amount of documentation work, should still be evaluated. As these questions remain open, a long road awaits before NLP methods can be considered successful in clinical care documentation.

3.2 Named entity recognition

Traditionally named entity recognition has focused on detecting proper nouns, often names of persons, organizations and locations (Mikheev et al., 1999), but the categories of interest have broadened over time to e.g. products, events and quantities (Luoma et al., 2020). Understandably the focus in the biomedical domain has been on entities relevant for molecular biology, microbiology or clinical studies, including entities such as genes, bacteria, diseases and treatments (Tanabe et al., 2005; Bossy et al., 2019; Pradhan et al., 2014; Uzuner et al., 2011).

Whereas named entity recognition itself is beneficial for certain applications, it is often combined with relation extraction (Bach and Badaskar, 2007) and named entity normalization (Khalid et al., 2008) for automated construction of knowledge bases (Van Landeghem et al., 2013).

In this section two different biomedical NER studies are addressed. The first one focuses on nursing narratives, similar to the data studied in Section 3.1.2, with interest in mentions of pain and acute confusion. The unique features of this study are the large amount of entity categories and relatively low amount of training examples for some of these classes. The study also discusses the relationship of named entity recognition and sentence-level text classification.

In the second study the aim is to identify pharmacological compounds from clinical case studies extracted from medical publications. This study is conducted in the scope of a shared task focusing on Spanish texts. Our emphasis in this study was on the assessment of minimally adapted multilingual and cross-lingual methods.

3.2.1 Related work

Conditional random fields (Lafferty et al., 2001) (see Section 2.1.5) have been the standard approach for named entity recognition for the past two decades although the feature representations have changed over time. Earlier methods have heavily relied on manually engineered features based on word and character level n-grams, other morphological features, part-of-speech tagging and shallow parsing (Nadeau and Sekine, 2007; Ritter et al., 2011). Another common information utilized has been existing ontologies and dictionaries (Mikheev et al., 1999; Kaewphan et al., 2017).

CRFs did not disappear during the emergence of neural transfer learning, but instead were often combined with CNN or LSTM layers (Ma and Hovy, 2016; Lample

et al., 2016). These methods no longer relied on feature engineering, but instead were based on *shallow* pretrained word embeddings.

A latter group of models relies on deeper transfer learning (Akbik et al., 2018) and some approaches have since stopped incorporating CRFs in the neural architectures (Devlin et al., 2019). These models are discussed further in Section 3.2.3.

Other recent studies have for instance focused on multitask learning (Chai et al., 2022), active learning (Shen et al., 2017) or cross-sentence and cross-document information (Luoma and Pyysalo, 2020; Wang et al., 2021).

3.2.2 Mentions of pain and acute confusion in Finnish medical documents

Pain management is a crucial part of the provided care as nearly all patients experience postoperational pain (Heikkilä et al., 2016). Similarly acute confusion is a common syndrome among geriatric patients (Voyer et al., 2008). Failures in pain management and identification of acute confusion have been associated with prolonged rehabilitation, yet the documentation of pain and delirium in patient care has been considered insufficient (Heikkilä et al., 2016; Voyer et al., 2008). Information extraction has the potential to enable large scale inspection of mentions of symptoms related to pain and confusion as well as their associations to varying procedures and situations as presented in care documentation.

Here we focus on the evaluation of named entity recognition methods for two Finnish medical datasets focusing on concepts related to pain and acute confusion. Named entity recognition by predicting the type and precise character offsets of a relevant phrase within the given text is the most common definition for NER and methods designed to model such structural task are usually employed. However, for many secondary medical documentation uses, such as information retrieval and text summarization, elaborate information extraction is not always necessary as the outcome is ultimately read by domain experts. Thus, we hypothesize that instead of detecting exact phrases expressing certain selected concepts, reducing the task complexity to sentence level classification could help in detecting more challenging concepts. To this end, in this study we do not only evaluate the performance of standard NER methods on entity-level, but also convert such predictions to sentence level labels and compare these to a direct sentence classifier.

Data

In this study, manually annotated datasets containing Finnish physician notes and nursing narratives are used. The documents are gathered from patients with an open heart surgery at any point in their medical history, i.e. although they cover a wide range of topics, a bias towards cardiovascular diseases can be expected. For both

pain and acute confusion related concepts independent annotation schemas have been formed, yet the manual annotations have been produced on the same set of documents. In total the dataset contains 1327 days of nursing narratives and 2156 notes written by physicians, resulting in 3483 documents in total. Although the documents are the same, we consider these two tasks independent and for instance the data has been randomly divided to training, development and test sets separately for each task.

Both of these datasets feature two properties which differentiate them from common biomedical and clinical NER datasets:

- **Large amount of classes:** whereas most NER datasets focus on couple of entity types of interest, the pain data contains 15 distinct labels covering explicit, implicit and potential pain mentions as well as attributes such as location, intensity and quality of pain and other pain related concepts. An example sentence can be seen in Figure 8. Similarly the acute confusion annotations cover a wide range of confusion and delirium related symptom categories, totalling in 37 different entity types. Although rare, similarly comprehensive annotations are also present in publicly available biomedical datasets such as CRAFT, GENIA and ShARe corpora (Bada et al., 2012; Kim et al., 2003; Kelly et al., 2014). The full lists of included entity types of this study are present in the supplementary materials of Paper IV.

- **Limited entity frequencies:** most classes in the datasets have very limited amount of training data compared to biomedical benchmark datasets. For instance the Biocreative VI Bio-ID track dataset (Arighi et al., 2017) contains on average almost 16K annotated entities per entity type, whereas our pain dataset consists of only 1.3K annotations per type, a tenth of the Biocreative dataset, yet still a fairly reasonable amount. The acute confusion data is another magnitude smaller with mere 110 annotations per entity type. Although these numbers are still somewhat hopeful, the entity type distributions are non-uniform: the rarest types contain only a handful of annotated occurrences. These rare entities include types such as *substance induced delirium* for acute confusion and *patient education* for pain annotations.

In addition to the annotated data, we use a large unstructured clinical text corpus with almost million physician notes and nursing narratives in total. This data is used for pretraining the word embeddings utilized in our methods.

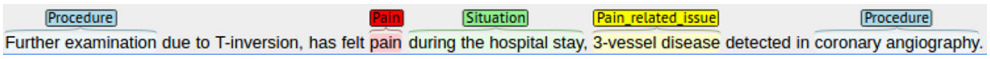

Figure 8. An example sentence with pain related entity annotations. Freely translated from the Finnish source material. Source: Paper IV

Methods

Three different approaches are developed and evaluated in this study. The first two methods model the task as sequence labelling: they predict a label for each token in the input sentence and also model the dependencies between neighbouring labels, i.e. the transitions from a label to another. Both of these methods rely on a CRF classifier, the first one, based on NERsuite [1], using more traditional features and the latter being a neural network with a CRF output layer. The third model is a sentence-level multilabel text classifier with the goal of predicting the existence of an entity within the sentence, i.e. the task is simplified from detecting to entity boundaries to only assessing whether an entity of a given type is present.

For the NERsuite based method two separate models are trained, one for each annotation schema. All input tokens are tagged with the IOB tagging schema, where the beginning and continuation tokens of the same entities are labeled with B and I respectively. As the data contains overlapping annotations, yet NERsuite does not support multilabel classification, we form combination classes for the overlapping instances. A common example is the Finnish compound word *rintakipu* (chest pain), which contains both *pain* and *location* entities. Notice that such issue rarely occurs in English, highlighting the importance of language technology studies focusing on other language groups as well.

NERsuite uses linguistic analyses such as part-of-speech tagging and lemmatization in feature generation. Because of this, it has been bundled with such tools, but the provided models are trained for English only. In this experiment we replace the preprocessing pipeline with the Finnish dependency parser (Haverinen et al., 2014), which provides both POS tagging and lemmatization, thus all used models in this study being trained on Finnish data. See Section 3.2.3 for cross-lingual NERsuite experiments.

As the second sequence labelling model, we use a CNN-BiLSTM-CRF model, following the approach by Ma and Hovy (2016). Similarly to the NERsuite it relies on a CRF output layer, but instead of requiring explicit feature generation and a complex preprocessing pipeline, it uses a CNN layer to analyze the input tokens as character sequences. The CNN-encoded representation of each token is then concatenated with a pretrained word embedding and the resulting vector is used as an input for a bidirectional LSTM layer. The LSTM encoding is subsequently used as the feature vector for the CRF and all layers are jointly trained. We use word2vec (Mikolov et al., 2013b) and the large clinical data corpus for pretraining the word embeddings. The target labels have been formed identically with the NERsuite models.

For the sentence classification approach, the final model consists of three independent LSTM blocks, each receiving either a sequence of tokens, lemmas or POS tags as the input, all represented with embeddings. The output for the last step in

[1] `http://nersuite.nlplab.org/`

Approach	Precision	Recall	F-score
	Pain		
NERsuite	87.29	62.88	73.10
CNN-BiLSTM-CRF	79.30	63.80	70.71
	Acute confusion		
NERsuite	69.33	36.84	48.11

Table 7. Mention-level evaluation of the tested NER approaches on the test sets of the Pain and Acute confusion corpora. The reported numbers are micro-averaged over the various classes.

the sequence is considered the encoding of the whole sequence and the output of all three LSTMs is concatenated and further fed to a fully connected output layer. Similarly to the CNN-BiLSTM-CRF model, the word embeddings are initialized with a word2vec model, but other embeddings are randomly initialized.

Results and Discussion

We firstly evaluate the NER models on entity level, using strict boundary matching criteria. Both NERsuite and CNN-BiLSTM-CRF models reach F-scores over 70% on pain related mentions (see Table 7). However, NERsuite shows slightly stronger performance and is used as the only NER model for acute confusion. Results on the latter dataset are much weaker with the best obtained F-score of only 48%. To compare the NERsuite results against the sentence classifier, we convert all detected entity mentions to sentence level predictions. These results suggest that surprisingly reducing the task complexity may not necessarily lead to stronger performance as the converted NERsuite predictions slightly outperform our best sentence classifier (Table 8). However, the overall scores for NERsuite on sentence classification are 5pp to 10pp higher than mention level scores, demonstrating that this metric is indeed relaxed compared to strict boundary matching.

Approach	Pain	Acute confusion
NER	78.61	59.41
SC	77.65	57.49

Table 8. Micro-averaged F-scores for the different approaches on the test sets of the pain and acute confusion data sets. NERsuite was used to produce the NER scores.

More detailed evaluation shows large performance variance across different entity types. For instance pain mentions (entity types *pain*, *implicit pain* and *potential pain*) seem to be relatively easy to detect with sentence level F-scores up to 96%. These are also fairly frequent entities and indeed a correlation between annotation counts and performance can be observed on this data (Pearson's R 0.55 for pain and

0.71 for acute confusion datasets). As the acute confusion dataset has far less annotations per entity type, the lack of training data has much larger impact on the performance, whereas the pain dataset already contains hundreds of annotations for most entity types and the benefit of additional training data starts to diminish.

Some entity types, such as *procedure* on the other hand are detected with far inferior performance although the amount of training annotations is high. This suggests that there might be less variance in certain classes like *pain* mentions, possibly due to the selected cohort of nursing documents. In fact, the ten most common unique *pain* mentions constitute over 40% of all pain mentions, whereas the ten most common *procedure* mentions cover only 23% of all such entities. Moreover, seven out of the ten most common *pain* mentions are inflectional forms of words ”kipu” (pain) or ”rintakipu” (chest pain). As our NERsuite model receives also the lemmatized variants of these tokens, it is apparent that the model will reach high performance, although word2vec should have also learnt the high semantic similarity of these inflectional forms.

To conclude, it seems that although neural approaches should benefit from transfer learning, particularly in low-resource settings such as the experiment described in this section, more traditional approaches may still reach equivalent or better performance. This may be partly due to the low variance in some of the entity types which leads to a situation where the models do not need the ability to generalize to semantically similar phrases. It is good to remind that these experiments were conducted during the time when shallow transfer learning in the form of non-contextualized word embeddings was still state-of-the-art approach and the task could be revisited with more recent methods.

3.2.3 Multi- and cross-lingual models in biomedical NER

A wide range of publicly available biomedical and clinical NER tools and datasets have been introduced for English (Settles, 2004; Leaman and Gonzalez, 2008; Weber et al., 2021; Doğan et al., 2014; Krallinger et al., 2015). Several of these originate from shared tasks such as BioCreative, SemEval and BioNLP Shared Task (Chen et al., 2021; Pradhan et al., 2014; Nédellec et al., 2013), which aid in comparison of various methods by establishing standardized evaluation platforms for the suggested systems. Unfortunately similar opportunities are rare for other languages. PharmaCoNER shared task of 2019 (Gonzalez-Agirre et al., 2019) is one of the few exceptions, focusing on extracting pharmacological compound mentions from Spanish clinical case studies included in medical publications. The main focus is on mentions of chemicals and proteins, including for instance antibodies and peptide hormones.

Our motivation for participating in this task was to access the effectiveness of applying multilingual and cross-lingual tools in clinical NER with minimal adaptation, as several languages such as Finnish lack both the needed training data as well

as publicly available tools for similar text mining purposes. The lack of training data could be greatly alleviated by using datasets created for other languages, if strong multilingual models could be trained. Unfortunately the lack of suitable evaluation data also makes it impossible for evaluating these multilingual approaches. Thus, we use the Spanish evaluation platform provided in PharmaCoNER for benchmarking.

We focus on two different approaches, the first one relying on an efficient biomedical NER tool, NERSuite, which however is only bundled with preprocessing models trained for English, and the second utilizing a multilingual BERT model. These approaches are described in greater detail in the following sections.

The data used in the PharmaCoNER shared task consists of 1,000 case studies covering a wide range of medical disciplines. The entities are divided into two main categories of chemicals (*Normalizables*) and proteins (*Proteinas*) with two smaller categories for chemical mentions that cannot be linked to existing ontologies (*No_normalizables*) and for general substance classes (*Unclear*). Overall the gold standard annotations contain 7,624 entity mentions of which over 97 percent belong to the two main categories.

Methods

Our first method is based on the NERSuite toolkit, a CRF-based classifier with a rich feature set tailored for English biomedical domain. For feature generation it relies on the GENIA tagger, which performs POS-tagging, lemmatization and chunking, but also utilizes features derived directly from the word forms. We use this tool in an off-the-shelf manner, i.e. we treat the Spanish data as English for feature generation purposes. The CRF classifier for NER tagging is still trained on the provided Spanish training data. Default hyperparameters are used.

The second tested model is based on a multilingual BERT model, trained on Wikipedia data collected for 104 languages (Pires et al., 2019). Although Spanish is one of the pretraining languages, the pretraining data is not of biomedical or clinical nature. Thus, this experiment is not only exploring the effectiveness of multilingual, but also cross-domain transfer learning. We use a cased variant of the multilingual BERT, which also preserves accented characters common in Spanish. As BERT relies on a vocabulary of subword units shared across all pretraining languages, it can benefit from commonalities of similar languages. On the other hand, the shared vocabulary leads to compromised subword representations due to varying uses in different languages and domains.

In our model, the pretraining output layers of the BERT model are removed and replaced with a CRF layer, which produces the final sequence of predicted labels. The selected hyperparameters closely follow the values mentioned in the original BERT paper Devlin et al. (2019), except for weight decay, which is optimized with a grid search. More details on the training approach are described in paper V. In the

Model	Precision	Recall	F-score
	Strict		
NERSuite	91.35	73.95	81.73
BERT	89.05	87.44	88.24
	Overlap		
NERSuite	95.45	77.37	85.47
BERT	93.71	91.85	92.77

Table 9. Evaluation of the proposed NER methods on the development data. *Strict* refers to an evaluation where mention offsets have to match exactly with the gold standard annotations, whereas in *Overlap* a predicted mention is considered correct as long as it overlaps with a gold standard mention, i.e. inaccuracies in beginning and end offsets are allowed.

prediction phase we always select the predicted label of the first subword unit as the label for the whole word.

Results and analysis

We use micro-averaged mention-level F-scores with strict boundary detection as our main evaluation metric, following the official PharmaCoNER evaluation. Overall results for the development set are shown in Table 9. Although NERSuite relies on English models for POS tagging, lemmatization and chunking, it still reaches an impressive performance with an F-score of 81.73%. However, the BERT model surpasses this baseline by a clear margin, with an F-score of 88.24%.

The entity type specific evaluation shows that the BERT model is consistently better than the NERSuite model with both of the main entity types *Proteinas* and *Normalizables*, the F-score being 6pp higher in both categories. However, the smaller categories of *Unclear* and *No_normalizables* demonstrate the superior generalization capability of the BERT model: +19pp difference to the NERSuite model. In fact, NERSuite is incapable of detecting any mentions of the smallest *No_normalizables* class, with only 24 training examples. Surprisingly, for the *Unclear* class with only 89 training examples, BERT still achieves an F-score of 82.11% on the development set, although this could be pure memorization of the specific mentions.

As our processing pipeline for BERT uses identical CoNLL input data format as NERSuite, the text is initially tokenized on word level and then further retokenized each token individually to subword units. For compatibility, the subword units are inversely detokenized back to the original format after prediction and the predicted label of the first subword unit of each word is assigned as the label for the whole word. The subword unit tokenization allows more freedom in predicting entity boundaries, which may cause additional offset errors and we thus use additional relaxed evaluation metrics to analyse the model performance. If errors are allowed in both beginning and end offsets of the detected mentions, as long as an overlap with a

Entities	Pretraining	No pretraining
All	87.44	54.00
Seen	96.13	70.16
Unseen	71.72	24.78

Table 10. Recall of the BERT model on development set with and without pretraining on all entities, entity spans which are also present in the training data (seen) and entity spans which do not appear in the training data (unseen).

gold standard entity is observed, both of the models show an improved performance (see Tabe 9). This suggests that the models are in fact able to detect more entities than is apparent from the strict evaluation, but create somewhat erroneous mention boundaries. However, the BERT model shows only slightly larger performance gap between strict and relaxed evaluations than the NERSuite model, implying that the additional subword tokenization steps do not drastically introduce additional entity boundary errors.

A key feature of using machine learning models in named entity recognition instead of plain dictionary matching is their ability to generalize to previously unseen entity mentions. For instance in the PharmaCoNER development dataset, 55% of the unique entity spans are not present in the training data. Thus, obtaining strong performance in this task requires either comprehensive dictionaries of pharmacological compounds or a model with strong generalization ability to new entity mentions and new contexts. For a harsh generalization estimate of the BERT model, we evaluate it on development set entities which are not present in the training data. The recall of the model on all entities, entities present in training data and entities not present in training data are listed in Table 10. Whereas the overall recall of the model is over 87%, it is virtually able to memorize all training entities with a recall of 96% on this subset. However, the recall of unseen mentions is also 72%, an indication of a relatively strong generalization ability.

In addition to being pretrained with a masked language model objective (see Section 2.3.1), the BERT architecture is also far more complex than its NERSuite counterpart, leading to a situation where it is unclear whether the stronger performance can be attributed to transfer learning or simply the model's higher capacity for modelling complex sentence structures. The importance of transfer learning can be further questioned due to the multilingual and out-of-domain nature of the pretraining data. To validate that the pretraining step is indeed influental on the outcome, we also train a separate model, architecturally identical to the multilingual BERT, but with random initialization of all model parameters. This model obtains an abysmal performance on the development set with an F-score of 56% (see Table 11). The weak performance is not due to a poor learning, but instead caused by massive overfitting as the model reaches almost 100% F-score on the training data. When

Pretraining	Precision	Recall	F-score
Yes	89.05	87.44	88.24
No	57.62	54.00	55.75

Table 11. Impact of BERT pretraining on the development set performance.

measured only on previously unseen entity mentions (see Table 10), the recall of this model is mere 25%, 47pp weaker than with pretraining. It is worth noting though, that this overfitting was practically impossible to avoid due to the model's memorization capacity. Interestingly the randomly initialized model also suffers from a low recall on previously seen entity mentions (26pp weaker than pretrained model), suggesting that the model is not simply memorizing the training data mentions, but instead their full context and is unable to detect these mentions in other contexts.

Comparison to other PharmaCoNER models

As this study is part of a shared task, we have a great opportunity to compare our methods to other successful approaches suggested by the other participants. In particular, three different approaches of interest are discussed in this section: a neural baseline provided by the shared task organizers (Gonzalez-Agirre et al., 2019), a BiLSTM-CRF model with FLAIR embeddings and BERT as utilized by other competitors.

Meanwhile the PharmaCoNER shared task was ongoing, the organizers published a neural approach trained and evaluated on the same data (Armengol-Estapé et al., 2019). The model relies on character- and token-level (Bi)LSTM layers for encoding the input sentences accompanied with more traditional features used in NER models. Two variants of this model were introduced as baseline by the shared task organizers, one with word embeddings pretrained with general domain data and the other with medical data. The general domain model achieved an F-score of 82% in the official evaluation, a score barely higher than our NERSuite model, which has no language specific pretraining or feature-engineering. The model with medical word embeddings shows slightly stronger performance with an F-score of 85%, yet still over 2pp below our BERT model, which uses no domain-adapted pretraining. This shows that the multilingual BERT has surprisingly strong performance out-of-the-box, without any additional features or hyperparameter tuning, whereas models with shallow pretraining steps tend to require more task-specific tailoring to perform well.

Another approach, generally considered as state-of-the-art in sequence tagging, FLAIR (Akbik et al., 2018, 2019) was also tested in the shared task by Stoeckel et al. (2019). Whereas the original FLAIR model relies on character level representations of the examined word, in this work a latter variant is used. This version of FLAIR pools the character-based encodings of a word from various contexts in-

stead of using only the encoding of the examined context. In addition standard word embeddings are concatenated with the character-based encodings as also suggested in the original FLAIR model. Moreover, the authors extend the FLAIR model with additional subword unit embeddings. In their experiments plain FLAIR model with exclusive character-level modelling reaches an F-score of 86%, but a variant with three additional word or subword level embeddings added, obtains an outstanding performance level of over 90% in F-score, which is a considerable improvement over our scores. Notice though that in their approach the embeddings are pretrained with both general domain and clinical Spanish data, depending on the embeddings, which provides stronger specialization for the language and task at hand, but requires a notable amount of additional work and experimenting, as a multitude of word embeddings were trained and tested, yet all of them did not result in improved performance. However, FLAIR in its basic form still shows performance comparable to BERT. This has also been demonstrated in other NER tasks in the original FLAIR publications, where FLAIR is shown to surpass BERT's performance on the English CoNLL NER dataset (Akbik et al., 2019).

The best performing system suggested in the shared task is also based on multilingual BERT (Xiong et al., 2019). Similar to the FLAIR approach by Stoeckel et al. (2019) they expand standard BERT with additional embeddings, in this case with character, POS and word form embeddings. A CRF layer is used for the final output similarly to our approach. This approach reached an F-score of 91%, slightly outperforming the FLAIR system, with their plain BERT baseline also surpassing an F-score of 90% on the development set, a 2pp improvement over our model. However, a questionable approach is used in their training pipeline: after optimizing the model against the development set, the training and development sets are combined for further training. This is a standard approach for models, which are trained until convergence or can be optimized with analytical approaches. As BERT is not regularized enough to prevent overfitting and as the development set is no longer available for validating the correct amount of training steps, which can be considered a hyperparameter of its own, this method provides no options for measuring how strongly the model is overfitting. In their paper Xiong et al. (2019) mention that the combined training is continued for five epoch, but provide no further justification for this number. After a short discussion with the authors during the EMNLP conference, it seems that the training amount has been chosen arbitrarily. Thus, there is a chance that the strong performance has been achieved by accident and similar results may be hard to reproduce on other datasets, where the five epochs of combined training already overfits the model. Note that other methods discussed here do not utilize the development data for training, which makes model comparison difficult.

Further demonstrating the effectiveness of BERT, Sun and Yang (2019) follow similar minimalistic approach as we do by training BERT without any modifications and replace even the CRF layer with a fully connected layer. However, they per-

form a thorough hyperparameter search for the model, resulting in slightly better performance than our approach. The most distinguishable difference between these approaches is that their batch size is 4 times larger than the one used in our experiments. Note that such a large batch would not have fit into the memory of the GPU used in our experiments leading to their model not being trainable on any consumer grade GPUs present at the time these experiments were conducted without gradient accumulation over batches. However, this is a strong indicator that a larger batch size should have a positive regulatory effect on fine-tuning BERT for this task. Another interesting experiment they conduct is fine-tuning the BioBERT model (Lee et al., 2019) for this task, although BioBERT is pretrained only on English biomedical data. This experiment is similar to our cross-lingual NERsuite approach, but surprisingly their BioBERT model achieves almost the same performance as the multilingual model on this task. This raises the question whether pharmacological entity detection and English-Spanish language pair are a good platform for evaluating multilingual approaches as these entity mentions are lexically very similar in both languages.

3.2.4 Summary and future work

The previous sections have summarized our work on medical text classification and named entity recognition. Our studies suggest that these tasks have a strong overlap, named entity recognition being structurally more challenging to model. However, Bertian models seem to narrow down this difference as the previously utilized conditional random fields are no longer necessary for strong performance, resulting in almost identical models being able to solve these two tasks.

Although our study shows that BERT's strong performance stems from the effective transfer from the pretraining tasks, it is unclear how much the model relies on non-contextualized word embeddings and how much the context is in fact utilized. A future work direction is to probe Transformer models (Bibal et al., 2022; Wu et al., 2020b) for better understanding of how much they benefit from the context in tasks similar to named entity recognition. This line of research would connect our interpretation study on the Finnish clinical text classification model latent space with the more recent Transformer-based approach used in the PharmaCoNER task.

Another direction is to explore how well the multilingual BERT performs on Finnish clinical data, although a Finnish BERT model (Virtanen et al., 2019) has also been released after the studies conducted in this thesis, as a multilingual model has intuitively more potential for cross-lingual transfer. For tasks such as the PharmaCoNER, similar datasets with slightly varying definitions often exists for other languages. For instance the CHEMDNER corpus (Krallinger et al., 2015) contains equivalent chemical and drug mentions as the PharmaCoNER, but in English. For Finnish the most interesting future task is to utilize such resources in a cross-lingual

zero-shot setting, as no similar datasets are available for Finnish fine-tuning of the models. Finnish and English being more distant languages would also establish a more challenging benchmarking platform for cross-lingual transfer than the English-Spanish language pair used in our studies.

3.3 Protein function prediction

Whereas high-throughput sequencing and mass-spectrometry have resulted in a vast amount of structural information on proteins, more detailed studies on their functional properties tend to be laborious (Zhou et al., 2019). Computational methods have thus been suggested as a scalable alternative for functional annotation of proteins.

In this study we explore an ensemble of machine learning models, in particular neural networks and random forests in the task of protein function prediction. The input in this task is a protein, in the form of an amino acid sequence and the goal is to suggest a set of functional properties the protein has. The possible functions, covering molecular functions, biological processes and cellular components, are based on the Gene Ontology (GO) definitions (Ashburner et al., 2000; The Gene Ontology Consortium, 2018) and form a closed set. Thus the problem can be addressed as a multilabel classification task with tens of thousands of possible categories.

Although this study seems disconnected from the themes of this thesis, from the machine learning perspective it is analogous with the text classification tasks discussed in previous sections: if treated as a sequence of amino acids a protein can be considered as a piece of text, each amino acid corresponding to a character or a word, with their own structural and functional roles in the sequence. An important standpoint in our study is thus to explore the possibilities of modelling the functional properties of a protein directly from the amino acid sequence, using neural network models similar to those discussed in Section 3.1. Due to the evident similarities between language and protein modelling, NLP methods, in particular the recent language models have become inspirational in the bioinformatics field (Ofer et al., 2021). The related and subsequent studies exploring such approaches are discussed in Section 3.3.4.

3.3.1 The Critical Assessment of Functional Annotation

Our study has been conducted in the context of the Critical Assessment of Functional Annotation (CAFA) shared task. CAFA provides an evaluation platform where the participants are requested to provide functional predictions for a large set of amino acid sequence. Following the prediction submissions, several months are waited for some of the target sequences to accumulate experimentally verified functional annotations. This subset is then used for the official evaluations.

The first version of the system described in this section was initially created for the third CAFA challenge organized in 2016-2017 and it placed in top-10 out of over 100 submissions. Small changes to the approach have been made since.

3.3.2 Methods

We train and evaluate three independent classifiers for the task: a feedforward neural network (FNN), a random forest (RF) and a convolutional neural network (CNN). The first two rely on manually crafted features whereas the CNN model focuses solely on the amino acid sequence.

To train the models we select a subset of proteins from the Swiss-Prot (The UniProt Consortium, 2020) protein sequence database. The subset is based on functional annotations with reliable experimental evidence or manual curation, resulting in over 67,000 proteins with nearly 400,000 functional annotations. As the GO annotations form a hierarchical structure, we also add the ancestral GO concepts to the datasets, increasing the number of individual annotations to almost 4,000,000. In our experiments the dataset is split into separate training, validation and test subsets and we focus only on the 5,000 most common GO concepts, which cover over 94 percent of all the annotations present in the data. On our evaluation, however, the full annotation set is used as gold standard, i.e. the missing GO concepts outside the 5,000 most common ones are automatically counted as false negative errors and our evaluation scores are not artificially inflated.

The manually curated features used can be categorized into four different groups. The first set of features describes the similarity of the protein to other proteins in the Swiss-Prot database and is based on the BLASTP sequence alignment search (Camacho et al., 2009). The second set of features utilizes the InterProScan (Jones et al., 2014) toolkit to provide structural and functional information of the proteins. Thirdly the taxonomical information of the organism of origin of the given protein is retrieved from the NCBI Taxonomy database (Federhen, 2012). The last feature group is based on tools such as NucPred (Heddad et al., 2004), NetAcet (Kiemer et al., 2005) and PredGPI (Pierleoni et al., 2008) to describe the localization and post-translational modifications of the analyzed proteins.

The experiments with the Random Forest classifiers were conducted by Jari Björne and their detailed description is thus left out of this thesis, as Kai Hakala did not contribute to this part of the study. However, it is worth mentioning that the random forest models were used for feature selection purposes and the optimal subset of features as evaluated on the validation set was also used as is with the FNN model.

The FNN is a standard fully connected neural network with a single hidden layer. As the output dimensionality corresponds to the 5,000 GO concepts to be predicted and the original number of features utilized is over 600,000, the hidden layer is re-

duced to the size of 300 nodes. In addition, another feature selection phase is conducted by removing features with variance below a certain threshold. We boost the recall of the model by modifying the standard binary cross entropy to penalize false negative predictions more than false positive ones. The magnitude of this penalty is an adjustable hyperparameter which is optimized on the validation dataset as a simple grid search.

Unlike the FNN and RF classifiers described above, the CNN model does not utilize any handcrafted features, but tries to predict the functional annotations directly from the amino acid sequence. The intuition behind the CNN approach is to learn fuzzy patterns in the amino acid sequences which correspond to certain structural or functional properties. Our model represents each unique amino acid as a latent embedding and a protein is consequently modelled as a sequence of these embeddings, analogously to sentences being modelled as sequences of token embeddings in the previous sections. As in language technology such embeddings have been traditionally initialized with a pretrained language model, we also experiment with initializing the amino acid embeddings with their physicochemical properties. These values are gathered from the Amino Acid Index database (Kawashima et al., 2007) and describe e.g. the hydrophobicity and flexibility of the amino acids. However, we did not observe any performance improvement compared to randomly initialized embeddings.

The embeddings are subsequently fed as an input to a convolutional layer. We use kernel sizes of 3, 9, 27 and 81. The narrow kernels correspond to common n-gram features and local segment lengths used in protein secondary structure prediction (Wang et al., 2016). The biological relevance of the wider kernels is that they should be able to detect whole motifs and shorter domains (Xu and Nussinov, 1998). For each size we learn 50 kernels. The kernel activations are then max pooled and used as an input to a fully connected output layer. It is possible that crucial location information is lost in the pooling process, but we have not experimented with alternative approaches so far.

3.3.3 Results

Our main evaluation metric is F-score, whereas the official CAFA evaluation uses the maximal F-score found on the precision-recall curve. Thus the CAFA evaluation does not force the systems to produce categorical decisions, whereas our evaluation is more strict and sets the lower bound for the CAFA evaluation metric.

The overall results are shown in Table 12. As a baseline we use a BLAST based homology transfer (HT) where known GO concepts of the most similar amino acid sequences found in the BLAST search are directly transferred to the targeted protein. A similar baseline is also used in the CAFA evaluations (Radivojac et al., 2013; Zhou et al., 2019).

NN	RF	HT	mode	F	P	R
CNN				0.347	0.316	0.385
FNN				**0.480**	0.492	0.468
	RF			0.424	0.609	0.326
		HT		0.387	0.550	0.298
FNN	RF		OR	**0.493**	0.472	0.517
FNN		HT	OR	0.487	0.460	0.518
	RF	HT	OR	0.471	0.527	0.426
FNN	RF		AND	0.398	0.707	0.277
FNN		HT	AND	0.363	0.675	0.248
	RF	HT	AND	0.312	0.740	0.198
FNN	RF	HT	OR	**0.493**	0.445	0.553
FNN	RF	HT	AND	0.296	0.765	0.184

Table 12. All ensemble combinations of the three methods: random forest classifier (RF), feedforward neural network (FNN), and the homology transfer (HT), using either intersection *AND* or union *OR* of the predictions. The performance is evaluated as the micro-averaged F-score (F), precision (P) and recall (R). Source: Paper VI

The FNN and RF classifiers outperform the baseline method with a clear margin, the neural model being the best performing classifier with an F-score of 0.48, almost 8pp improvement over the random forest model. As these models utilize the same features and random forests are generally considered strong baseline classifiers, this is a surprisingly large difference in performance. Partially this can be explained by the poor balance between precision and recall in the RF model, which was intentionally solved with the modified loss function of the neural model. The CNN model lacks in performance, being slightly weaker than the homology transfer baseline and over 13pp worse than the FNN model.

In addition to the individual models we form simple ensemble methods by taking the union and intersection of the predictions. The best overall performance is achieved by taking the union of the FNN and RF predictions, although this only leads to a minor improvement of 1.3pp over the FNN model.

To obtain a deeper understanding of the performance, we examine the results on the different GO main categories of molecular functions, biological processes and cellular components separately. We also look into the depth of these hierarchical ontologies and different groups of organisms.

For molecular functions, biological processes and cellular components our best performing ensemble results are 0.536, 0.446 and 0.620 in F-score. This seems to follow the general trend in the CAFA evaluations as the results for other methods are also much lower for the biological processes (Zhou et al., 2019). It is important to note that the biological process ontology has the highest average depth of leaf con-

cepts, i.e. the ontology seems to be the most detailed one, which intuitively explains the difficulty of predicting these concepts accurately. Overall our model submitted to the third CAFA challenge was in the top 10 best performing systems in all three categories. The main difference between our CAFA model and the model described in this thesis is that the CNN and FNN components were combined with a shared output layer in the shared task, a design decision later on identified as detrimental for the performance.

Looking into the most studied organism in Swiss-Prot reveals that the number of annotated proteins does not necessarily correlate with a better performance. Instead it seems that our model has particularly strong performance for prokaryotes (see Figure 9). The official CAFA evaluation confirms this as our ensemble model placed in top 3 for molecular function and cellular component categories of prokaryotic proteins. As most of the CAFA test set proteins are from eukaryotic origins, the overall performance is not heavily influenced by the prokaryotes.

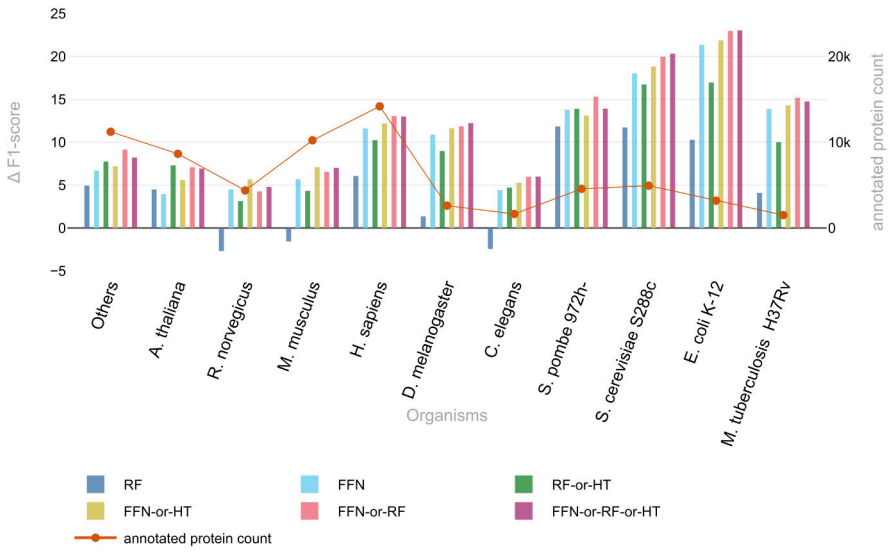

Figure 9. The performance of the methods on the 10 organisms that have more than 1,000 annotated protein sequences in the training data. *Others* represents a group of organisms with less than 1,000 annotated protein sequences. The vertical bars plot the difference of each tested method to the Homology Transfer fallback (left vertical scale). The red connecting line plots the number of annotated proteins for each organism (right vertical scale). Source: Paper VI

3.3.4 Discussion

Although the methods introduced in our study show competitive performance compared to other models evaluated in the third CAFA challenge, the overall performance of automated protein function prediction is still lacking. Moreover, it seems that only minor improvements have been made since the second iteration of CAFA (Jiang et al., 2016; Zhou et al., 2019).

As most of the introduced models still rely on manually crafted features, searching for new sources of information is critical. For instance some evidence of the benefits of co-expression and binding data has already been presented by Piovesan et al. (2015) and Kulmanov et al. (2018).

Another alternative is to find more powerful ways of utilizing the amino acid sequences directly. Our CNN model, although still weak, demonstrates that current neural models are able to learn meaningful patterns present in the protein sequences and implicitly model their structure and function. In itself, such a CNN model is already a noteworthy alternative for BLAST-based homology transfer, as the inference speed is considerably faster than a BLAST search. However, it seems that the small Swiss-Prot subset of high quality annotations is not a sufficient dataset for training more complex models. An important future direction is to assess the benefits of using the lower quality predicted annotations as silver standard data or the possibilities of self-supervised pretraining akin to language models. Whereas Bertian models have become popular also in the analysis of proteins (Nambiar et al., 2020; Qiao et al., 2021; Brandes et al., 2022) and strong results have been demonstrated in the closely related protein structure prediction task (AlQuraishi, 2019; Senior et al., 2020), the breakthrough in protein function prediction is yet to become.

4 Conclusions

This thesis focuses on the applications of neural networks in clinical and biomedical text mining. Here we highlight the main findings of each study included in this thesis. More task-specific and technical conclusions and insights have been discussed separately for text classification, named entity recognition and protein function prediction at the end of each corresponding section, namely Section 3.1, Section 3.2 and Section 3.3.

Papers I and II discuss sentence-level text classification of nursing notes into thematic categories, a task currently done manually. Our results show that neural text classification methods are able to achieve performance comparable to domain experts in this task, suggesting that such methods could already be deployed in clinical environments and thus reduce the documentation workload of nurses. In this study we have not, however, tried to determine how such a system should be optimally utilized in production nor how the impact of such a system should be measured. I should add though that measuring the improvements in clinical information systems has been widely studied and remains a challenging task, incorporating metrics for e.g. information quality, user satisfaction and productivity (Van Der Meijden et al., 2003; Nguyen et al., 2014). Moreover, the long-term impact of an EHR implementation may be different from the initial results (Baumann et al., 2018). In our case, the interesting questions are whether using the classification model leads to better information quality due to the more unified sentence labels and due to the added freedom of documentation. However, it is also possible that completely free narratives are slower to write, resulting in more time spent in documentation even if the nurses do not have to provide the thematic labels for the sentences written. An early study on electronic health records also shows that reducing the documentation workload does not necessarily result in more resources for actual care, but the time made available can be adversely spent on other difficulties introduced by the new system if not carefully designed (Allard et al., 1995).

The strong performance of the methods also suggests that the used approach of utilizing paragraph headings as weak sentence-level labels and framing the task as a multilabel classification problem are a valid strategy although the labels are not actually mutually exclusive. However, using such data for automated evaluation remains problematic and makes interpreting the results more challenging. Currently reiterating the experiments with more modern machine learning models would most likely

require another round of laborious manual evaluation and thus as a future work the automated evaluation setting should be improved, e.g. by taking into consideration the hierarchical nature and relatedness of the labels.

We also show preliminary results on a secondary use of the trained model in the analysis of the used clinical ontology based on the learned concept representations of the model. The results suggest that the currently used FinCC ontology is too complex for the nurses to use efficiently in care documentation. However, this part of the study should be further investigated and reflected against the existing studies on ontology evaluation (Brewster et al., 2004; Brank et al., 2005; McDaniel and Storey, 2019) and ontology learning (Ayadi et al., 2019; Shroff et al., 2021), although it seems similar approaches have not been previously studied. In general, it is clear that studies on such secondary uses of neural language technology models are far and few between, yet deserve more attention.

The Paper IV similarly focuses on Finnish nursing notes, but the aim is to detect relevant information on entity-level. In this task we did not observe any benefits in using transfer learning based neural methods, albeit the low-resource nature of the task with only a handful of annotated examples for the rare classes should be an excellent candidate for transfer learning techniques. It is good to keep in mind though that these experiments were conducted with the previous generation of shallow transfer learning instead of the recent Transformer models. Another main observation made in this study is that although named entity recognition is structurally more difficult task than mere text classification, simplifying the task from the former to the latter may not result in better results. If entity-level annotations are available for a given task, it thus seems more favourable to solve the task in this form and derive e.g. sentence-level categorization as a post-processing step instead if such labelling is needed in the downstream tasks.

In Paper III we design neural text classification methods for social media content. This paper focuses more on the methodological side of text classification with neural model designs incorporating transferred pretraining information, more classic bag-of-words features as well as grammatical features. Controlling the variance of random initialization in neural network training is also discussed. Retrospectively, it must be said that most of the findings of this study are already obsolete and may not apply to more modern neural methods.

Unlike the other studies in this thesis, Paper V utilizes a Transformer based approach and can be said to be still somewhat current method-wise. In this study we show that the recent language models with deep transfer can be extremely efficient in a cross-domain transfer learning setting, outperforming shallow transfer methods even when they have been trained on language- and domain-specific data.

Whereas Paper VI diverges from the natural language processing domain, this study utilizes the same neural models as discussed in the other papers and treats amino acid sequences similarly to sequences of text. For each unique amino acid

an embedding is learned and a CNN-based model is subsequently utilized for detecting functionally relevant patterns in the sequence. Although this approach alone did not result in state-of-the-art results, it is on par with commonly used BLAST-based homology searches. This is a strong indicator that such a model could replace BLAST-based database searches in certain applications, bypassing the need to maintain large protein sequence indices and reducing the computational cost of the search. A clear flaw in our approach is that the network was not pretrained on large unlabelled protein sequence data, although such resources are openly available. Instead the amino acid embeddings were initialized with features describing their physicochemical properties and the rest of the model was randomly initialized. It seems that the amount of labelled data used in this task is not enough for learning the task without any manual feature crafting. Notice that in Paper V a comparison of identical network architectures with and without pretraining was conducted for text analysis and the results suggest that pretraining is by far the most valuable asset in training complex neural models with limited finetuning data. In contrary, in Papers I and II we had access to millions of training sentences and it is unlikely that reiterating these studies with more recent pretraining methods would lead to considerably better results.

Insights on original research objectives and future work

The original research objectives of this thesis revolved around the applicability of modern neural networks in clinical and biomedical text analysis. In the papers included in the thesis, all three commonly used neural architecture categories have been considered: recurrent, convolutional and self-attentive. As these models were extensively evaluated on varying datasets, it can be claimed that the first research objective of developing and evaluating neural methods for clinical text classification and NER has been met.

This work was done during a technological cataclysm, in which self-attentive models marginalized all other model architectures. It seems though that the superiority of this design is often overstated and the other architectures are still relevant in many cases. For instance Tay et al. (2021) suggest that similarly pretrained CNNs can outperform their Transformer counterparts. Convolutional and recurrent components have also been combined with self-attention e.g. in Conformers (Gulati et al., 2020) and Block-recurrent transformers (Hutchins et al., 2022). In the studies included in this thesis similar trend can be seen: all tested architectures perform roughly the same, if pretrained and fine-tuned identically. Overall, it seems that finding optimal pretraining tasks and data is far more important than the choice of the model architecture. Unfortunately directly comparable large-scale pretraining runs of various different model configurations require computational resources only available to the largest companies in the field.

When it comes to the applicability of neural models in clinical text mining, it is clear that the models are performing well enough to be practical. In papers I and II we demonstrate that even the previous generation CNN and RNN models reach the performance level of domain experts if enough training data is available. Similarly in the scope of the PharmaCoNER shared task (Gonzalez-Agirre et al., 2019), on which the Paper V focuses on, the best performing systems are merely 2pp below inter-annotator agreement in performance. I conclude that the performance of modern language technology methods is not an obstacle for utilizing such tools in clinical care environments.

More important future question seems to be how these models should actually be deployed in production, taking into consideration the impact on the workflow of clinicians, as poorly designed systems have a high risk of reducing productivity. It is also unclear what are the main goals in applying language technology in clinical environments, e.g. patient safety, standardization of documentation practices, reduction of documentation time or more complex secondary documentation uses. Productivity in this context can also be measured objectively, as the time it takes to complete a task, or from the perspective of perceived cognitive workload (Wilbanks and McMullan, 2018). These goals can be contradicting as it is clear, that minimizing the completion time of the documentation task may lead to higher cognitive workloads and ultimately to an increase in medical errors. Moreover, measuring the impact of new technological advancements in clinical environments should be assessed over long periods of time as the influence of technological maturity and existing organizational processes on the adaptation of new technologies is still poorly understood (Malm-Nicolaisen et al., 2022).

For low-resource languages multilingual and cross-domain approaches are critical and one of the goals of this thesis was to explore these methods in the clinical domain. However in our studies we only scratched the surface on this topic with the experiments described in Paper V. In this work focusing on Spanish case studies, we were able to demonstrate strong results with multilingual and cross-domain knowledge transfer. However, I also suspect that the close similarity of the biomedical Spanish and English languages simplifies this task so drastically, that it cannot be considered a good platform for deriving conclusions for how well the tested cross-lingual and multilingual methods would work for other languages. I have to emphasize again, that studies focusing on low-resource languages are still important as the evidence for cross-lingual transfer in biomedical domain is too scarce. Unfortunately for Finnish no multilingual models were tested in this thesis, leaving this line of research behind as a future work.

The fourth objective of this thesis focused on the interpretation and utilization of the neural language representations for purposes the models were not originally trained for. My only study conducted for this line of work is the utilization of the trained neural text classifier for the evaluation of the clinical ontology usage in Pa-

per II. As this was a very preliminary study, it is hard to draw any general conclusions from this experiment. Yet, it demonstrates that such secondary uses are possible and should be further studied.

The last research objective was to evaluate how well protein function prediction can be treated as a text classification task with amino acid sequences as the input sequence. The results suggest, that protein function prediction is much harder task albeit the structural similarity and the current methods for modelling such sequential data are not efficient enough.

Although in this thesis only the small sub-problem of finding the correct subject headings for clinical care documentation sentences and paragraphs has been studied, the overall goal is to fully automate the documentation process. In the long run this seems feasible with the combination of speech recognition, language technology and sensory data. For instance the instructions provided to the patient could be documented directly by analyzing the dialogue between the patient and the care providers. Similarly medical procedures are also often explained to the patient during the care. Preliminary studies on this topic already exist, but are often limited to simple consultation conversations between the patients and physicians (Molenaar et al., 2020; Yim and Yetisgen, 2021). The emerging multimodal large language models enable combining visual and other sensory data with the dialogue audio (Driess et al., 2023; Huang et al., 2023), although I do not see such models being utilized in this context in the near future. Again, how these tools are used in practice should be carefully scrutinized as they may influence the patient-clinician interaction and subsequently the medical outcomes (Kelley et al., 2014). In the optimal case the tools might even steer the dialogue to a more informative direction, benefiting also the patient.

While writing the conclusions of this thesis, the generative large language models have been causing massive hype, probably even more so than the earlier Transformer-based models such as Bert a few years before. The instruction and dialogue fine-tuned versions of such models enable a new form of extracting information from these models via prompting the models with natural language questions or instructions (Ouyang et al., 2022). As the outcome of this type of information extraction is highly dependent on the form of the used prompts, these models have created a new line of work named *prompt engineering*. No standard practices in academic research or commercial utilization of this technique have been established so far, although the exploitation of prompting is currently heavily studied (Brown et al., 2020; Wei et al., 2021, 2022; Liu and Chilton, 2022; White et al., 2023; Zhou et al., 2022; Zamfirescu-Pereira et al., 2023). Although prompting is shown to be an effective zero-shot approach, it is brittle (Wei et al., 2022) and model dependent (White et al., 2023) leading to poorly understood and hard-to-maintain systems. Enforcing interpretability and explicit structures in the latent representations of language models might prevent language technology from turning into black magic.

Another major problem, in particular for low-resource languages, is the data-

efficiency of large language models. For instance the MassiveText dataset (Rae et al., 2021) contains over 2 trillion tokens worth of text, out of which a subset of couple of hundred billion tokens is often used for training a language model. For comparison, the language exposure of young children is estimated to be up to 14 million words per year (Montag et al., 2018). Thus the language exposure of a language model trained on 300B tokens, e.g. Gopher and GPT-3, corresponds roughly to 20,000 years of human life, yet these models do not demonstrate general intelligence similar to humans. Very promising results have been achieved on computational efficiency (Hu et al., 2021) and parameter-efficiency (Zhang et al., 2023), but with data-efficiency I have not seen similar results. Although language technology has advanced drastically during the past years, the pretraining learning objectives are effectively the same as three decades ago and similarly, the earliest gradient descent based optimization methods have been introduced already in the 19th century (Lemaréchal, 2012), with no fundamental differences to the methods currently in use. Novel learning paradigms are essential if we aim to narrow the gap between language models and human language acquisition.

Bibliography

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 724–728.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Allard, J., Dzwonczyk, R., Yablok, D., Block Jr, F., and McDonald, J. (1995). Effect of automatic record keeping on vigilance and record keeping time. *BJA: British Journal of Anaesthesia*, 74(5):619–626.

AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics*, 35(22):4862–4865.

Arighi, C., Hirschman, L., Lemberger, T., Bayer, S., Liechti, R., Comeau, D., and Wu, C. (2017). Bio-ID track overview. In *Proc. BioCreative Workshop*, volume 482, page 376.

Armengol-Estapé, J., Soares, F., Marimon, M., and Krallinger, M. (2019). PharmacoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics Inform*, 17(2):e15–.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

Ayadi, A., Samet, A., de Beuvron, F. d. B., and Zanni-Merk, C. (2019). Ontology population with deep learning-based NLP: a case study on the Biomolecular Network Ontology. *Procedia Computer Science*, 159:572–581.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bach, N. and Badaskar, S. (2007). A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., et al. (2012). Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):1–20.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baumann, L. A., Baker, J., and Elshaug, A. G. (2018). The impact of electronic health record systems on clinical documentation times: A systematic review. *Health policy*, 122(8):827–836.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., Francois, T., and Watrin, P. (2022). Is attention explanation? An introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bossy, R., Deléger, L., Chaix, E., Ba, M., and Nédellec, C. (2019). Bacteria biotope at BioNLP open shared tasks 2019. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 121–131.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Brank, J., Grobelnik, M., and Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170. Citeseer Ljubljana Slovenia.

Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1):1–9.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chai, Z., Jin, H., Shi, S., Zhan, S., Zhuo, L., and Yang, Y. (2022). Hierarchical shared transfer learning for biomedical named entity recognition. *BMC bioinformatics*, 23(1):1–14.

Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Chalmers, I. and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, 374(9683):86–89.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.

Chen, Q., Allot, A., Leaman, R., Doğan, R. I., and Lu, Z. (2021). Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas. Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

de Gibert, O., Perez, N., Garcia-Pablos, A., and Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Duque, A., Martinez-Romo, J., and Araujo, L. (2016). Can multilinguality improve biomedical word sense disambiguation? *Journal of biomedical informatics*, 64:320–332.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Friedrichs, J., Mahata, D., and Gupta, S. (2018). InfyNLP at SMM4H task 2: Stacked ensemble of shallow convolutional neural networks for identifying personal medication intake from twitter. *arXiv preprint arXiv:1803.07718*.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D., and Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, 383(9913):267–276.

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., and Krallinger, M. (2019). PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. pages 5036–5040.

Hakala, K. (2015). UTU: Adapting biomedical event extraction system to disorder attribute detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 375–379. Association for Computational Linguistics.

Hakala*, K., Kaewphan*, S., Björne*, J., Mehryary, F., Moen, H., Tolvanen, M., Salakoski, T., and Ginter, F. (2020). Neural network and random forest models in protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Hakala, K., Kaewphan, S., Salakoski, T., and Ginter, F. (2016). Syntactic analyses and named entity recognition for PubMed and PubMed Central —up-to-the-minute. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pages 102–107. Association for Computational Linguistics.

Hakala, K., Mehryary, F., Kaewphan, S., and Ginter, F. (2013a). Hypothesis generation in large-scale event networks. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM'13)*, pages 19–28.

Hakala*, K., Mehryary*, F., Moen, H., Kaewphan, S., Salakoski, T., and Ginter, F. (2017). Ensemble of convolutional neural networks for medicine intake recognition in Twitter. In *SMM4H@ AMIA*, pages 59–63.

Hakala, K. and Pyysalo, S. (2019). Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61.

Hakala, K., Van Landeghem, S., Kaewphan, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). CyEVEX: Literature-scale network integration and visualization through cytoscape. In *Proceedings of SMBM'12, Zurich, Switzerland*, pages 91–96.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2013b). EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop (BioNLP-ST'13)*, pages 26–34.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2015). Application of the EVEX resource to event extraction and network construction: Shared Task entry and result analysis. *BMC Bioinformatics*, 16(Suppl 16):S3.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Hall, A. and Walton, G. (2004). Information overload within the health care system: a literature review. *Health Information & Libraries Journal*, 21(2):102–108.

Hanka, R. and Fuka, K. (2000). Information overload and "just-in-time" knowledge. *The Electronic Library*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Haverinen, K., Ginter, F., Laippala, V., and Salakoski, T. (2009). Parsing clinical Finnish: Experiments with rule-based and statistical dependency parsers. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 65–72.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531.

He, P., Liu, X., Gao, J., and Chen, W. (2021). DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Heddad, A., Brameier, M., and MacCallum, R. M. (2004). Evolving regular expression-based sequence classifiers for protein nuclear localisation. In *Workshops on Applications of Evolutionary Computation*, pages 31–40. Springer.

Heikkilä, K., Peltonen, L.-M., and Salanterä, S. (2016). Postoperative pain documentation in a hospital setting: A topical review. *Scandinavian journal of pain*, 11(1):77–89.

Hendrich, A., Chow, M. P., Skierczynski, B. A., and Lu, Z. (2008). A 36-hospital time and motion study: how do medical-surgical nurses spend their time? *The Permanente Journal*, 12(3):25.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

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:S362–S368.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. (2023). Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Hughes, M., Li, I., Kotoulas, S., and Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Stud Health Technol Inform*, 235:246–50.

Hunt, R. E. and Newman, R. G. (1997). Medical knowledge overload: a disturbing trend for physicians. *Health care management review*, 22(1):70–75.

Hutchins, D., Schlag, I., Wu, Y., Dyer, E., and Neyshabur, B. (2022). Block-recurrent transformers. *arXiv preprint arXiv:2203.07852*.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers), pages 1681–1691. Association for Computational Linguistics.

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):1–19.

Joachims, T. et al. (1999). Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Kaewphan, S., Hakala, K., and Ginter, F. (2014). UTU: Disease mention recognition and normalization with CRFs and vector space representations. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 807–811. Association for Computational Linguistics and Dublin City University.

Kaewphan, S., Hakala, K., Miekka, N., Salakoski, T., and Ginter, F. (2018). Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. *Database*, 2018.

Kaewphan, S., Mehryary, F., Hakala, K., Salakoski, T., and Ginter, F. (2017). TurkuNLP entry for interactive Bio-ID assignment. In *Proceedings of the BioCreative VI Workshop*, pages 32–35. Bethesda, MD, USA.

Kanerva, P., Kristoferson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22.

Kavuluru, R., Rios, A., and Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic acids research*, 36(suppl_1):D202–D205.

Kelley, J. M., Kraft-Todd, G., Schapira, L., Kossowsky, J., and Riess, H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. *PloS one*, 9(4):e94207.

Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D. L., Velupillai, S., Chapman, W. W., Martinez, D., Zuccon, G., et al. (2014). Overview of the ShARe/CLEF ehealth evaluation lab 2014. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 172–191. Springer.

Khalid, M. A., Jijkoun, V., and De Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

Kiemer, L., Bendtsen, J. D., and Blom, N. (2005). NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, 21(7):1269–1270.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Klein, A., Sarker, A., Rouhizadeh, M., O'Connor, K., and Gonzalez, G. (2017a). Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system. In *BioNLP 2017*, pages 136–142.

Klein, A., Sarker, A., Rouhizadeh, M., O'Connor, K., and Gonzalez, G. (2017b). Detecting personal medication intake in Twitter: Annotation guidelines.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kulmanov, M., Khan, M. A., and Hoehndorf, R. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Larsen, P. and Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Leaman, R. and Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008*, pages 652–663. World Scientific.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Lemaréchal, C. (2012). Cauchy and the gradient method. *Doc Math Extra*, 251(254):10.

Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140. IEEE.

Lin, H.-C. and Chang, C.-M. (2018). What motivates health information exchange in social media? The roles of the social cognitive theory and perceived interactivity. *Information & Management*, 55(6):771–780.

Liu, F., Weng, F., and Jiang, X. (2012). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1035–1044.

Liu, P., Qiu, X., and Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2873–2879. AAAI Press.

Liu, V. and Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., and Pyysalo, S. (2020). A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4615–4624, Marseille, France. European Language Resources Association.

Luoma, J. and Pyysalo, S. (2020). Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 904–914, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Malm-Nicolaisen, K., Fagerlund, A. J., and Pedersen, R. (2022). How do users of modern EHR perceive the usability, user resistance and productivity five years or more after implementation? *Studies in health technology and informatics*, 290:829–833.

McDaniel, M. and Storey, V. C. (2019). Evaluating domain ontologies: clarification, classification, and challenges. *ACM Computing Surveys (CSUR)*, 52(4):1–44.

McInnes, B. T., Pedersen, T., and Pakhomov, S. V. (2009). UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In *AMIA Annual Symposium Proceedings*, volume 2009, page 431. American Medical Informatics Association.

Mehryary, F., Hakala, K., Kaewphan, S., Björne, J., Salakoski, T., and Ginter, F. (2017). End-to-end system for bacteria habitat extraction. In *Proceedings of the 2017 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

Mehryary, F., Kaewphan, S., Hakala, K., and Ginter, F. (2014). Eliminating incorrect events from large-scale event networks by trigger word clustering and pruning. In *Proceedings of SMBM'14*, pages 75–79.

Mehryary, F., Kaewphan, S., Hakala, K., and Ginter, F. (2016). Filtering large-scale event collections using a combination of supervised and unsupervised learning for event trigger classification. *Journal of Biomedical Semantics*, 7(1):1–13.

Melas-Kyriazi, L. (2021). Do you even need attention? A stack of feed-forward layers does surprisingly well on ImageNet. *arXiv preprint arXiv:2105.02723*.

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., and Lehmann, C. U. (2017). Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics*, 26(1):38.

Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Moen*, H., Hakala*, K., Mehryary, F., Peltonen, L.-M., Salakoski, T., Ginter, F., and Salanterä, S. (2017). Detecting mentions of pain and acute confusion in Finnish clinical text. In *BioNLP 2017*, pages 365–372.

Moen*, H., Hakala*, K., Peltonen, L.-M., Matinolli, H.-M., Suhonen, H., Terho, K., Danielsson-Ojala, R., Valta, M., Ginter, F., Salakoski, T., et al. (2020a). Assisting nurses in care documentation: from automated sentence classification to coherent document structures with subject headings. *Journal of Biomedical Semantics*, 11(1):1–12.

Moen*, H., Hakala*, K., Peltonen, L.-M., Suhonen, H., Ginter, F., Salakoski, T., and Salanterä, S. (2020b). Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods. *Journal of the American Medical Informatics Association*, 27(1):81–88.

Moen, H., Hakala, K., Peltonen, L.-M., Suhonen, H., Loukasmäki, P., Salakoski, T., Ginter, F., and Salanterä, S. (2018). Evaluation of a prototype system that automatically assigns subject headings to nursing narratives using recurrent neural network. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 94–100.

Molenaar, S., Maas, L., Burriel, V., Dalpiaz, F., and Brinkkemper, S. (2020). Medical dialogue summarization for automated reporting in healthcare. In *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 76–88. Springer.

Momenipour, A. and Pennathur, P. R. (2019). Balancing documentation and direct patient care activities: A study of a mature electronic health record system. *International journal of industrial ergonomics*, 72:338–346.

Montag, J. L., Jones, M. N., and Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive science*, 42:375–412.

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., and Hoving, C. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4):e85.

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., Shaikh, K., and Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert systems with applications*, 116:494–520.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Icml*.

Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). Transforming the language of life: Transformer neural networks for protein prediction tasks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, New York, NY, USA. Association for Computing Machinery.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., and Zweigenbaum, P. (2013). Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 1–7.

Nguyen, L., Bellucci, E., and Nguyen, L. T. (2014). Electronic health records implementation: an evaluation of information system impact and contingency factors. *International journal of medical informatics*, 83(11):779–796.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Ofer, D., Brandes, N., and Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Paul, M. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. B., and Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In *Biological,*

translational, and clinical language processing, pages 97–104, Prague, Czech Republic. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pierleoni, A., Martelli, P. L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC bioinformatics*, 9(1):1–11.

Piovesan, D., Giollo, M., Ferrari, C., and Tosatto, S. C. (2015). Protein function prediction using guilty by association from interaction networks. *Amino Acids*, 47(12):2583–2592.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Pradhan, S., Chapman, W., Man, S., and Savova, G. (2014). Semeval-2014 task 7: Analysis of clinical text. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014*. Citeseer.

Pyysalo, S., Campos, J., Cejuela, J. M., Ginter, F., Hakala, K., Li, C., Stenetorp, P., and Jensen, L. J. (2015). Sharing annotations better: RESTful Open Annotation. In *Proceedings of ACL'15: Demonstrations*.

Qiao, Y., Zhu, X., and Gong, H. (2021). BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*, 38(3):648–654.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.

Robinson, K. E. and Kersey, J. A. (2018). Novel electronic health record (EHR) education intervention in large healthcare organization improves quality, efficiency, time, and impact on burnout. *Medicine*, 97(38).

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pages 901–909.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

Sarker, A., Maksim, B., Jasper, F., Hakala, K., Svetlana, K., Mehryary, F., Han, S., Tran, T., Rios, A., Kavuluru, R., de Bruijn, B., Ginter, F., Mahata, D., Mohammad, S. M., Nenadic, G., and Gonzalez-Hernandez, G. (2018). Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*.

Schenk, E., Schleyer, R., Jones, C. R., Fincham, S., Daratha, K. B., and Monsen, K. A. (2018). Impact of adoption of a comprehensive electronic health record on nursing work and caring efficacy. *CIN: Computers, Informatics, Nursing*, 36(7):331–339.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., and Anandkumar, A. (2017). Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Shi, Y., Karatzoglou, A., Baltrunas, L., Larson, M., Oliver, N., and Hanjalic, A. (2012). CLiMF: Learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, page 139–146, New York, NY, USA. Association for Computing Machinery.

Shroff, N., Vandenbussche, P.-Y., Moore, V., and Groth, P. (2021). Supporting ontology maintenance with contextual word embeddings and maximum mean discrepancy. In *DeepOntoNLP/X-SENTIMENT@ ESWC*, pages 11–19.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stoeckel, M., Hemati, W., and Mehler, A. (2019). When specialization helps: Using pooled contextualized embeddings to detect chemical and biomedical entities in Spanish. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 11–15, Hong Kong, China. Association for Computational Linguistics.

Sun, C. and Yang, Z. (2019). Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., et al. (2021). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

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 *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 720–727. Springer.

Suominen, H. J. and Salakoski, T. I. (2010). Supporting communication and decision making in Finnish intensive care with language technology. *Journal of Healthcare Engineering*, 1(4):595–614.

Sutskever, I., Martens, J., and Hinton, G. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1017–1024, Madison, WI, USA. Omnipress.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):1–7.

Tay, Y., Dehghani, M., Gupta, J. P., Aribandi, V., Bahri, D., Qin, Z., and Metzler, D. (2021). Are pretrained convolutions better than pretrained transformers? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4349–4359, Online. Association for Computational Linguistics.

The Gene Ontology Consortium (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338.

The UniProt Consortium (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A. P., Keysers, D., Uszkoreit, J., et al. (2021). MLP-Mixer: An all-MLP architecture for vision. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Uronen, L., Salanterä, S., Hakala, K., Hartiala, J., and Moen, H. (2022). Combining supervised and unsupervised named entity recognition to detect psychosocial risk factors in occupational health checks. *International Journal of Medical Informatics*, 160:104695.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Van Der Meijden, M., Tange, H. J., Troost, J., and Hasman, A. (2003). Determinants of success of inpatient clinical information systems: a literature review. *Journal of the American Medical Informatics Association*, 10(3):235–243.

Van Landeghem, S., Björne, J., Wei, C.-H., Hakala, K., Pyysalo, S., Ananiadou, S., Kao, H.-Y., Lu, Z., Salakoski, T., Van de Peer, Y., et al. (2013). Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4):e55814.

Van Landeghem, S., Hakala, K., Rönnqvist, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics, special issue Literature-Mining Solutions for Life Science Research*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Voorhees, E. M. et al. (1999). The TREC-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Voyer, P., Cole, M. G., McCusker, J., St-Jacques, S., and Laplante, J. (2008). Accuracy of nurse documentation of delirium symptoms in medical charts. *International Journal of Nursing Practice*, 14(2):165–177.

Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., and Singh, S. (2019). AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6(1):1–11.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., and Tu, K. (2021). Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., and Akbik, A. (2021). HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Wilbanks, B. A. and McMullan, S. P. (2018). A review of measuring the cognitive workload of electronic health records. *CIN: Computers, Informatics, Nursing*, 36(12):579–588.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020a). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Wu, Z., Chen, Y., Kao, B., and Liu, Q. (2020b). Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Xiong, Y., Shen, Y., Huang, Y., Chen, S., Tang, B., Wang, X., Chen, Q., Yan, J., and Zhou, Y. (2019). A deep learning-based system for PharmaCoNER. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 33–37, Hong Kong, China. Association for Computational Linguistics.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Xu, D. and Nussinov, R. (1998). Favorable domain size in proteins. *Folding and Design*, 3(1):11–17.

Yee, T., Needleman, J., Pearson, M., Parkerton, P., Parkerton, M., and Wolstein, J. (2012). The influence of integrated electronic medical records and computerized nursing notes on nurses' time spent in documentation. *CIN: Computers, Informatics, Nursing*, 30(6):287–292.

Yim, W.-w. and Yetisgen, M. (2021). Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., and King, I. (2018). Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3120–3131, Brussels, Belgium. Association for Computational Linguistics.

Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., and Qiao, Y. (2023). Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*.

Zheng, K., Vydiswaran, V. V., Liu, Y., Wang, Y., Stubbs, A., Uzuner, Ö., Gururaj, A. E., Bayer, S., Aberdeen, J., Rumshisky, A., et al. (2015). Ease of adoption of clinical natural language processing software: an evaluation of five systems. *Journal of biomedical informatics*, 58:S189–S196.

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2024

ISBN 978-951-29-9725-1 (PRINT)
ISBN 978-951-29-9726-8 (PDF)
ISSN 2736-9390 (Print)
ISSN 2736-9684 (Online)