MICHEL OLEYNIK

LEVERAGING WORD EMBEDDINGS FOR BIOMEDICAL NATURAL LANGUAGE PROCESSING

Dissertation

LEVERAGING WORD EMBEDDINGS FOR BIOMEDICAL NATURAL LANGUAGE PROCESSING

submitted by
Michel Oleynik

for the Academic Degree of
Doctor of Philosophy (PhD)

at the
Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation

under the Supervision of
Univ.-Prof. Dr.med. Stefan Schulz
Univ.-Prof. Dipl.-Ing. Dr. techn. Andrea Berghold
Univ.-Prof. Dr. Udo Hahn

2020

To my family.

DECLARATION

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organizations that have contributed to the research for this thesis. Due acknowledgment has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the "Standards of Good Scientific Practice and Ombuds Committee at the Medical University of Graz".

Graz, June 2020

Michel Oleynik

DISCLOSURES

COPYRIGHT STATEMENTS

Contents of the publication

Oleynik M, Kugic A, Kasác Z, and Kreuzthaler M. Evaluating Shallow and Deep Learning Strategies for the 2018 n2c2 Shared Task on Clinical Text Classification. JAMIA 2019;26:1247–54

were partially reused in Chapter 5 according to the Oxford Press Publication Rights[1]:

> The right to include the article in full or in part in a thesis or dissertation, provided that this is not published commercially.

Such reuse does not require obtaining prior permission:

> For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance.

Furthermore, the article is distributed under the terms of the CC-BY-NC license:

> This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

All co-authors are affiliated with the Institute for Medical Informatics, Statistics and Documentation of the Medical University of Graz in Austria. Markus Kreuzthaler is additionally affiliated with CBmed GmbH — Center for Biomarker Research in Medicine in Graz, Austria. All co-authors have explicitly agreed to the use of the published data in this thesis.

ON THE USAGE OF PERSONAL PRONOUNS

Several style guides in English oppose the use of passive voice in academic writing for better clarity; when using active voice though,

1 https://academic.oup.com/journals/pages/access_purchase/rights_and_permissions/publication_rights

accountability to the author must be given. In this thesis, I have chosen to explicitly differ my contributions (stated with first-person singular pronouns, e. g., "I" and "my") from ideas discussed with a group of people (stated with first-person plural pronouns, e. g., "we" and "our"). I believe such treatment provides the best balance between text clarity and accountability.

Chapter 4 had contributions from Johannes Hellrich and Stefan Schulz. Chapter 5 had contributions from Amila Kugic, Markus Kreuzthaler, and Zdenko Kasáč. Chapter 6 had contributions from Ariane Morassi Sasso, Erik Faessler, Jan Philipp Sachs, Pablo López-García, Stefan Schulz, and Zdenko Kasáč.

PUBLICATIONS

Though not always directly related, the following publications have been completed over the course of the PhD:

1. Oleynik M, Finger M, and Patrão DFC. Automated Classification of Pathology Reports. In: *MEDINFO 2015: eHealth-enabled Health - Proceedings of the 15th World Congress on Health and Biomedical Informatics, São Paulo, Brazil, 19-23 August 2015*. Ed. by Sarkar IN, Georgiou A, and Azevedo Marques PM de. Vol. 216. Studies in Health Technology and Informatics. IOS Press, 2015:1040.
2. Kreuzthaler M, Oleynik M, Avian A, and Schulz S. Unsupervised Abbreviation Detection in Clinical Narratives. In: *Proceedings of the Clinical Natural Language Processing Workshop, ClinicalNLP@COLING 2016, Osaka, Japan, December 11, 2016*. Ed. by Rumshisky A, Roberts K, Bethard S, and Naumann T. The COLING 2016 Organizing Committee, 2016:91–8.
3. Oleynik M, Patrão DFC, and Finger M. Automated Classification of Semi-Structured Pathology Reports into ICD-O using SVM in Portuguese. In: *MIE 2017: Informatics for Health: Connected Citizen-Led Wellness and Population Health, Manchester, United Kingdom, 24-26 April 2017*. Vol. 235. Studies in Health Technology and Informatics. IOS Press, 2017:256–60.
4. Oleynik M, Kreuzthaler M, and Schulz S. Unsupervised Abbreviation Expansion in Clinical Narratives. In: *MEDINFO 2017: Precision Healthcare through Informatics - Proceedings of the 16th World Congress on Medical and Health Informatics, Hangzhou, China, 21-25 August 2017*. Ed. by Gundlapalli AV, Jaulent M, and Zhao D. Vol. 245. Studies in Health Technology and Informatics. IOS Press, 2017:539–43.
5. López-García P, Oleynik M, Kasác Z, and Schulz S. TREC 2017 Precision Medicine - Medical University of Graz. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*. Ed. by Voorhees EM and Ellis A. Vol. 500-324. NIST Special Publication. National Institute of Standards and Technology (NIST), 2017.
6. Oleynik M, López-García P, Kasáč Z, and Schulz S. A FOSS Framework for Ranking Biomedical Documents at TREC-PM. Medical Informatics Europe (MIE 2018). 2018.
7. López-García P, Oleynik M, Kasáč Z, and Schulz S. Information Retrieval for Precision Oncology. ÖPPM - Gemeinsame Impulse für Personalisierte Medizin. 2018.

8. Oleynik M, Faessler E, Sasso AM, Kappattanavar A, Bergner B, Cruz HFD, et al. HPI-DHC at TREC 2018 Precision Medicine Track. In: *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*. Ed. by Voorhees EM and Ellis A. Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology (NIST), 2018.
9. Kreuzthaler M, Oleynik M, Vera-Ramos JA, Kasáč Z, and Schulz S. Secondary Use of Clinical Problem List Entries for Data-Driven Learning Approaches. In: *Workshop Booklet of the 10th International Workshop on Simulation and Statistics, Salzburg, Austria, 2-6 September 2019*. 2019:62.
10. Oleynik M, Kugic A, Kasác Z, and Kreuzthaler M. Evaluating Shallow and Deep Learning Strategies for the 2018 n2c2 Shared Task on Clinical Text Classification. JAMIA 2019;26:1247–54.
11. Faessler E, Hahn U, and Oleynik M. JULIE Lab & Med Uni Graz @ TREC 2019 Precision Medicine Track. In: *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*. Ed. by Voorhees EM and Ellis A. Vol. 1250. NIST Special Publication. National Institute of Standards and Technology (NIST), 2019.
12. Kreuzthaler M, Oleynik M, and Schulz S. Character-Level Neural Language Modelling in the Clinical Domain (to appear). In: *MIE 2020: Proceedings of the 30th Medical Informatics Europe*. Studies in Health Technology and Informatics. IOS Press, 2020.
13. Pomares-Quimbaya A, López-Úbeda P, Oleynik M, and Schulz S. Leveraging PubMed to create a Specialty-Based Sense Inventory for Spanish Acronym Resolution (to appear). In: *MIE 2020: Proceedings of the 30th Medical Informatics Europe*. Studies in Health Technology and Informatics. IOS Press, 2020.
14. Hahn U and Oleynik M. Medical Information Extraction in the Age of Deep Learning (to appear). Yearb. Med. Inform. 2020.
15. Faessler E, Oleynik M, and Hahn U. What makes a Top-Performing Precision Medicine Search Engine? Tracing Main System Features in a Systematic Way (to appear). In: *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Xi'an, China, July 25-30, 2020*. ACM, 2020.

Though this be madness, yet there is method in't.

— William Shakespeare (1602)

To infinity... and beyond.

— Buzz Lightyear (1995)

ACKNOWLEDGMENTS

A PhD is a long — sometimes arduous, sometimes joyful — journey through intense personal and professional growth. It was a pleasure to share this journey with so many incredible people, without whom it would not be possible to bring this work to an end.

First, I thank my supervisor, Stefan Schulz, for his guidance throughout these years. Stefan did not only actively engage in (sometimes endless!) brainstorming sessions but also supported me on a personal level several times. He also gave me the freedom I needed to pursue my curiosity in whichever direction I wanted. I also thank the members of my dissertation committee, Andrea Berghold and Udo Hahn, for the support and the valuable feedback.

I would like to express my gratitude to all members of the Institute for Medical Informatics, Statistics, and Documentation for their overall support. I am grateful to the Brazilian National Council for Scientific and Technological Development (CNPq) for funding my research (grant 206892/2014-4) during the four initial years and the Medical University of Graz through the PhD Program on Advanced Medical Biomarker Research (AMBRA).

In specific, I thank my colleagues who worked daily with me and participated in several spontaneous discussions on science and life. The Spanish clan brought a bit of Latin culture to my daily life: Catalina Martínez-Costa, José Antonio Miñarro Giménez, José Antonio Vera Ramos, Pablo López-García. Zdenko Kasáč, thanks for all the assistance and out-of-the-box conversations. Special thanks go to Markus Kreuzthaler for the afternoon brainstorming sessions and Stefanie Jauk for her kind friendship and unconditional support.

I also want to thank my colleagues from the Digital Health Center at the Hasso Plattner Institute (Potsdam, Germany). Special thanks go to Ariane Morassi Sasso for all the hard work (including proofreading the full thesis) and sharing her life in Berlin when an energy boost was needed.

Similarly, I thank my colleagues from the JULIE Lab at the Jena University (Jena, Germany) for welcoming me during my research stay in June 2018. A big "thank you" must be devoted to Christina

Lohr for connecting me to the team and to Erik Faessler for the endless help and amazing discussions — I learned a lot from you.

Moving abroad did not automatically teach me how to deal with homesickness. I thank my parents, João Oleynik and Maria Ester Helpa Oleynik, for investing in my education and encouraging my decisions even when that implied staying far away. I am incredibly proud of my brother, Leonardo Oleynik, who inspires me to be the older brother he deserves.

I found nonetheless an amazing family in Graz that enormously helped turn this city into a new *Heimat*. Marina Muñoz, Juan Carlos Hurtado Sierra, Beatrice Nanni, Matteo Saya, Peggy Ang Pei Yee: thank you for all the incredible moments we had together.

Last, but not least, I am grateful to my life partner Thisby Alarcón Khury Oleynik for embarking with me on this journey — thank you.

CONTENTS

1 INTRODUCTION 1
- 1.1 Motivation 1
- 1.2 Goals and Objectives 3
- 1.3 Outline 3

2 BACKGROUND 5
- 2.1 Machine Learning 5
- 2.2 Transfer Learning 5
- 2.3 Natural Language Processing 7
- 2.4 Word Sense Disambiguation 9
- 2.5 Text Classification 10
- 2.6 Information Retrieval 11

3 MATERIALS AND METHODS 13
- 3.1 Materials 13
- 3.2 Evaluation Metrics 14
- 3.3 Methods 16

4 CLINICAL TEXT CLEANSING 19
- 4.1 Introduction 19
- 4.2 Materials 23
- 4.3 Methods 26
- 4.4 Results 28
- 4.5 Discussion 30

5 CLINICAL TEXT CLASSIFICATION 35
- 5.1 Introduction 35
- 5.2 Materials 38
- 5.3 Methods 41
- 5.4 Results 43
- 5.5 Discussion 47

6 BIOMEDICAL INFORMATION RETRIEVAL 53
- 6.1 Introduction 53
- 6.2 Materials 56
- 6.3 Methods 59
- 6.4 Results 64
- 6.5 Discussion 68

7 CONCLUSION 73
- 7.1 Synopsis 73
- 7.2 Findings 73
- 7.3 Outlook 74

A APPENDIX 77
- A.1 Clinical Text Classification 77
- A.2 Biomedical Information Retrieval 77

BIBLIOGRAPHY 83
INDEX 101

ACRONYMS

BA	Biomedical Abstract
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
CT	Clinical Trial
DL	Deep Learning
EHR	Electronic Health Record
ICD	International Classification of Diseases
idf	inverse document frequency
infNDCG	inferred Normalized Discounted Cumulative Gain
IR	Information Retrieval
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NDCG	Normalized Discounted Cumulative Gain
n2c2	National NLP Clinical Challenges
NB	Naive Bayes
NCBI	National Center for Biotechnology Information
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
NLP	Natural Language Processing
NN	Neural Network
PMC	PubMed Central
RBC	Rule-Based Classifier
RNN	Recurrent Neural Network
SVM	Support Vector Machine
tf	term frequency
tf-idf	term frequency - inverse document frequency
TREC	Text REtrieval Conference
TREC-PM	Text REtrieval Conference - Precision Medicine
UMLS	Unified Medical Language System
WE	Word Embeddings

LIST OF FIGURES

Figure 4.1 2D PCA representation of word embeddings 28
Figure 4.2 F_1-score at different ranks 31
Figure 5.1 A sample n2c2 document 39
Figure 5.2 Overall F_1 score per criterion on the test set 44
Figure 5.3 Overall accuracy per criterion on the test set 45
Figure 6.1 Overview of the TREC-PM task 57
Figure 6.2 A sample TREC-PM topic 57
Figure 6.3 Overview of the TREC-PM assessment process 58
Figure 6.4 TREC-PM framework 60
Figure 6.5 BA: overall infNDCG, P@10, and recall 65
Figure 6.6 BA: P/R curves 65
Figure 6.7 CT: overall NDCG, P@10, and recall 67
Figure 6.8 CT: P/R curves 67
Figure A.1 BA: NDCG, P@10, and recall per topic 81
Figure A.2 CT: NDCG, P@10, and recall per topic 82

LIST OF TABLES

Table 3.1 Public datasets used in this thesis 14
Table 3.2 Pre-trained word embeddings 14
Table 3.3 Contingency table 15
Table 4.1 Related work on acronym resolution 21
Table 4.2 Acronym expansion dataset characteristics 24
Table 4.3 Acronym expansion annotation guidelines 25
Table 4.4 Counts of acronym types 29
Table 4.5 Acronym expansion results 30
Table 4.6 WORD2VEC expansions (examples) 32
Table 5.1 n2c2 dataset splits 38
Table 5.2 Classification criteria overview 40
Table 5.3 Overview of the proposed methods 41
Table 5.4 Overall results on the test set 46
Table 5.5 n2c2 top overall systems 50
Table 6.1 TREC-PM data overview 58
Table 6.2 Positive and negative keyword boosters 62
Table 6.3 BIO.NLPLAB.ORG expansions (examples) 63
Table 6.4 Gene family names (examples) 63
Table 6.5 BA: top overall systems 71
Table 6.6 CT: top overall systems 71

Table A.1 Official results: Baseline 77
Table A.2 Official results: Rule-based classifier 78
Table A.3 Official results: SVM 78
Table A.4 Official results: SELF-LR 79
Table A.5 Official results: PRE-LR 79
Table A.6 Official results: SELF-LSTM 80
Table A.7 Official results: PRE-LSTM 80

ABSTRACT

Driven by the decreasing costs of whole genome sequencing, the field of Precision Medicine has gained traction to allow targeted treatment choices for patients with specific biomarkers. To reach that goal, automated processing of the huge unstructured data pool in electronic health records and online resources such as PubMed is necessary to aid time-constrained health professionals not only in delivering precise treatments but also in building representative cohorts for new clinical trials. While Natural Language Processing (NLP) has shown great progress in several domains with the public availability of huge collections that enable Deep Learning (DL) approaches, the same has not been seen in the clinical field due to ethical concerns with data and model sharing. To overcome that disparity, recent advances in transfer learning methods — such as context-based Word Embeddings (WE) — have facilitated partial reuse of these large models.

With the overall goal of improving precision medicine, this work leverages WE for biomedical NLP in three lines of research: (a) clinical text cleansing; (b) clinical text classification; and (c) biomedical information retrieval. In the line of research (a), I demonstrated a novel method that associates WE with a minimal set of filtering rules to expand acronyms in a totally unsupervised way. This method outperformed traditional approaches using both n-grams and a hand-crafted sense inventory. In the line of research (b), I explored several methods for clinical phenotyping and cohort building. I verified that logistic regression associated with WE constituted a better model for clinical text classification than more complex DL architectures and also determined that embeddings pre-trained on a larger corpus were not better than embeddings trained on the target dataset. Finally, in the line of research (c), we proposed a method for query expansion that does not affect the precision of results in a biomedical information retrieval scenario. Using this method, I showed that WE could be effectively used to increase recall when structured resources were not available and additionally revealed that the benefit of query expansion was larger in a small dataset.

ZUSAMMENFASSUNG

Aufgrund der abnehmenden Kosten der Genomsequenzierung hat das Feld der Präzisionsmedizin an klinischer Bedeutung gewonnen, indem gezielte Behandlungen für PatientInnen mit spezifischen Biomarkern ermöglicht werden. Um dieses Ziel zu erreichen, ist die automatisierte Verarbeitung großer, unstrukturierter Datensätze in elektronischen Gesundheitsakten und Online-Ressourcen wie PubMed notwendig. Damit können die unter Zeitdruck stehenden MitarbeiterInnen des Gesundheitssystems bei individuellen Behandlungen unterstützt und der Aufbau repräsentativer Kohorten für neue klinische Studien erleichtert werden. Während Natural Language Processing (NLP) in vielen Feldern aufgrund der öffentlichen Verfügbarkeit umfangreicher Datensätze große Fortschritte durch Deep Learning (DL)-Ansätze erzielen konnte, sind solche Anwendungen im klinischen Bereich aufgrund ethischer Bedenken bezüglich des Datenschutzes eher selten. Um dieser Problematik entgegenzuwirken, wurde durch jüngste Fortschritte in den Methoden des Transfer-Learning — wie z. B. durch kontextbasierte Word Embeddings (WE) — die teilweise Wiederverwendung großer Modelle erleichtert.

Mit dem Ziel, die Präzisionsmedizin weiter zu verbessern, werden in dieser Arbeit WE für klinische NLP in folgenden drei Forschungsbereichen eingesetzt: (a) Bereinigung von klinischen Texten; (b) Klassifizierung von klinischen Texten; und (c) biomedizinische Informationssuche. Im Forschungsbereich (a) habe ich eine neue Methode demonstriert, die WE mit einem minimalen Satz von Filterregeln verbindet, um Akronyme vollständig unüberwacht auf ihre Langform abzubilden. Diese Methode übertraf die traditionellen Ansätze, welche sowohl n-Gramme als auch ein manuell erstelltes Inventar an Wortbedeutungen verwenden. Im Rahmen des Forschungsbereichs (b) habe ich mehrere Methoden zur klinischen Phänotypisierung und Kohortenbildung untersucht. Dabei fand ich heraus, dass die mit WE kombinierte, logistische Regression ein besseres Modell für die Klassifikation von klinischen Texten darstellte als komplexere DL-Architekturen. Zudem habe ich gezeigt, dass Embeddings, die auf einem größeren Datensatz vortrainiert wurden, keine besseren Ergebnisse zeigten als Embeddings, die auf dem spezifischen Zieldatensatz trainiert wurden. Für den Forschungsbereich (c) haben wir eine Methode zur Anfrageerweiterung entwickelt, welche die Präzision der Ergebnisse in einem biomedizinischen Informationssuche-Szenario nicht beeinträchtigt. Mit dieser Methode konnte ich zeigen, dass WE die Trefferquote erhöhen können, wenn keine strukturierten Ressourcen verfügbar waren, und dass der Nutzen der Anfrageerweiterung für kleinere Datensätze größer war.

RESUMO

Impulsionado pelos custos decrescentes de sequenciamento do genoma, o campo da Medicina de Precisão vem ganhando força por permitir escolhas de tratamento direcionadas para pacientes com biomarcadores específicos. Para atingir esse objetivo, é necessário processar automaticamente enormes coleções de dados não estruturados em registros eletrônicos de saúde e recursos on-line, como PubMed, para ajudar os profissionais de saúde com tempo limitado não apenas no fornecimento de tratamentos precisos, mas também na construção de coortes representativas para novos ensaios clínicos. Embora técnicas de Natural Language Processing (NLP) venham apresentando um grande progresso em várias áreas com a disponibilidade pública de grandes coleções que permitem técnicas de Deep Learning (DL), o mesmo não tem sido observado no domínio clínico devido a preocupações éticas com o compartilhamento de dados e modelos. Para superar essa disparidade, recentes avanços nos métodos de transfer learning — como Word Embeddings (WE) baseadas em contexto — facilitaram a reutilização parcial desses grandes modelos.

Com o objetivo geral de melhorar a Medicina de Precisão, esse trabalho utiliza WE para NLP clínica em três linhas de pesquisa: a) limpeza de textos clínicos; (b) classificação de textos clínicos; e (c) recuperação de informações biomédicas. Na linha de pesquisa (a), demonstrei um novo método que associa WE a um conjunto mínimo de regras de filtragem para expandir acrônimos de uma maneira totalmente não-supervisionada. Esse método superou as abordagens tradicionais usando n-gramas e um inventário de significados manual. Na linha de pesquisa (b), explorei vários métodos para fenotipagem clínica e construção de coorte. Verifiquei que regressão logística associada a WE constituiu um modelo melhor para classificação de texto clínico do que arquiteturas DL mais complexas e também determinei que embeddings pré-treinadas em uma coleção maior não foram melhores do que embeddings treinadas no conjunto de dados de destino. Finalmente, na linha de pesquisa (c), propusemos um método para expansão de buscas que não afeta a precisão dos resultados em um cenário de recuperação de informações biomédicas. Usando esse método, mostrei que WE puderam ser usadas para aumentar a cobertura quando recursos estruturados não estavam disponíveis e também revelei que o benefício da expansão da busca foi mais significativo em um conjunto de dados menor.

1 INTRODUCTION

1.1 MOTIVATION

Precision medicine

The field of Precision Medicine has gained traction by allowing targeted treatment choices [2], mainly driven by the decreasing costs of whole genome sequencing. It builds upon the growing knowledge that some therapies are more effective on patients presenting specific characteristics, including molecular (e. g., genomic, proteomic, metabolomic), clinical (e. g., age, sex, comorbidities), lifestyle, and environmental — e. g., blood transfusions are guided by blood typing for over hundred years [3, 4]. Analyzing these characteristics together in novel studies — e. g., Genome-Wide Association Studies (GWAS) — contributes to identifying patient profiles that present a given trait.

Precision Medicine relies therefore on a good match between (a) data that characterize an individual patient and (b) medical knowledge about diagnosis and care, not only regarding concluded but also ongoing research projects. In the first case (a), physicians have to process increasingly large amounts of data from a variety of information sources in Electronic Health Record (EHR) systems, such as images, biosignals, omics, texts, numeric, and coded data. In the second case (b), physicians would typically consult literature databases like PubMed for relevant literature in the field. However, in cases where the person is critically ill and standardized treatment options do not exist yet (or when existing treatments have proven ineffective), physicians may also consult clinical trials portals such as ClinicalTrials.gov for new therapeutic options to follow (given previous informed consent) or, depending on the availability of the clinical study, directly enroll the patient [5, 6].

However, given the time constraints clinicians are commonly subject to, continuous automated matching of patient data with biomedical research may be necessary to aid health professionals in effectively delivering precise treatments at the point of care [7]. Additionally, the mass processing of real-world clinical data (so-called secondary data use) is crucial not only to identify trends, but also to build representative cohorts for clinical trials that improve statistical power during result analysis [8, 9].

Structured clinical data issues

Automated processing is facilitated by the availability of structured data. Structured clinical data is nonetheless commonly subject to data quality issues — such as sparsity, bias, uncertainty, and incompleteness [10] — that hinder its direct reuse. For instance, while in the intensive care unit biosignals are collected at a high sampling rate, in normal

care laboratory values are recorded only on demand and data may thus be sparse. Wherever coded data is created for a different purpose than scientific analysis (e. g., billing [11]), bias can be expected; additionally, the fact that some data were collected may represent a bias in itself, since data collection may follow a preliminary (maybe undocumented) diagnostic hypothesis. Data may also be recorded at a different time point than when they were collected and thus be uncertain. Finally, structured data may simply be missing or not available at the required level of detail.

Clinical text issues

As structured clinical information may not be informative enough, unstructured sources such as text (e. g., discharge summaries) may be used as an alternative [12]. So-called Natural Language Processing (NLP) comes however with its own challenges to deal with syntax and semantics [13]. Moreover, clinical texts usually display a myriad of specific linguistic phenomena (e. g., misspellings, abbreviations, acronyms, and short/incomplete sentences) that gives them a telegraphic style [14] — sometimes hard even for humans to understand. For instance, computer algorithms should be able to understand the relationship among word parts (e. g., that "colitis" is an inflammation of the colon), words in a sentence (e. g., that "denies fever" indicates an absence of fever), words and their senses (e. g., that "ECG" is a short form for "electrocardiogram"), and sentences in the overall document (e. g., that "calcium" has a different meaning in a medication list than in lab results). These characteristics make automated processing of clinical text even more difficult because traditional NLP models trained in the general domain cannot be directly reused in clinical texts.

NLP has nonetheless shown great progress in the general domain [15] with the public availability of huge (newspaper-style) collections that allow the use of complex Machine Learning (ML)[1] approaches — so-called Deep Learning (DL) — in an end-to-end fashion. However, the same has not been seen in the clinical domain because de-identified clinical data (i. e., content from health records) are mostly not publicly available due to ethical concerns [16]. This disparity has been addressed by recent advances in transfer learning methods that allow reusing models trained on large datasets on smaller collections (see Section 2.2). For instance, context-based word representations — so-called word embeddings — can be learned in a large corpus and used as input for ML models in a smaller domain.

With the overall goal of improving information processing in precision medicine, I hereby propose leveraging word embeddings for biomedical NLP in three distinct (but interconnected) lines of research: (a) clinical text cleansing; (b) clinical text classification; and (c) biomedical information retrieval. In the first line of research (a), I explore unsupervised methods to expand short forms such as abbreviations and acronyms (typically found in clinical documents) and thus im-

1 See Section 2.1 for an introduction on the topic.

prove further automated processing. In the second line of research (b), I explore techniques to categorize clinical text that can be used not only for clinical phenotyping but also for cohort building. Finally, in the last line of research (c), I evaluate the impact of several strategies to improve the retrieval of documents from large biomedical collections.

1.2 GOALS AND OBJECTIVES

1.2.1 *General Goal*

To *evaluate* word embeddings for biomedical natural language processing, namely text cleansing, text classification, and information retrieval.

1.2.2 *Specific Objectives*

- To *assess* word embeddings pre-trained in large corpora for transfer learning.
- To *compare* shallow and deep machine learning approaches for text classification.
- To *measure* the impact of domain knowledge expressed as rules on model efficiency.

1.3 OUTLINE

This work is structured as follows.

In Chapter 2, I present a background of the main scientific aspects of this work. I provide a brief introduction into machine learning techniques (Section 2.1), transfer learning methods (Section 2.2), and a short review of the state of the art of natural language processing (Section 2.3). I then introduce notions of word sense disambiguation (Section 2.4), text classification (Section 2.5), and information retrieval (Section 2.6) necessary to understand the core chapters of this thesis.

In Chapter 3, I give a description of the data used for this work (Section 3.1), the evaluation metrics used to compare models (Section 3.2), and methods to validate the results obtained thereof (Section 3.3).

I then focus on the main research questions of this thesis. In Chapter 4, I explore word embeddings associated with filtering rules for acronym expansion, an important step in clinical text cleansing. In Chapter 5, I study shallow and deep learning models for text classification, associated with pre-trained word embeddings and embeddings trained on the target corpus. In Chapter 6, I examine word embeddings for query expansion in information retrieval and compares them to traditional approaches based on terminologies and rules.

Finally, in Chapter 7, I summarize the contributions of this thesis (Section 7.1), revisit findings (Section 7.2), and provide an outlook on future work to be explored by other researchers (Section 7.3).

2 BACKGROUND

2.1 MACHINE LEARNING

Machine Learning (ML), a subarea of Artificial Intelligence, investigates methodologies and develops systems that optimize a criterion function based on a set of examples or past experience [17]. It is based upon the hypothesis that data collections have hidden patterns that can be inferred by a computer and then summarized on a simpler model.

Unsupervised machine learning

Supervised machine learning

The area is further divided into *unsupervised* and *supervised* learning. In the first case, learning happens on a dataset without output information, i. e., each example is just a point x, $x \in \mathbb{R}^n$. Therefore, the goal here is to find regular patterns in the example set by using, e. g., clustering techniques. In the second case, a set of examples is tagged with the output information, i. e., each example is a pair $(x, f(x))$, with $x \in \mathbb{R}^n$. The goal is then to train a model on the examples to be applied to new inputs — unknown to the system, but similar to the training examples.

Neural network

Inspired by neural structures in animal brains, artificial Neural Networks (NNs) have also been explored over decades for both unsupervised and supervised learning. Such networks are composed of layers of nodes (the neurons) connected by edges; the output of any given node is determined by an *activation function* (e. g., sigmoid, hyperbolic tangent, softmax, rectified linear unit) over the (weighted) inputs; weights are then iteratively learned using *backpropagation* [18], an efficient process to compute the gradient of the loss function with respect to the network weights.

Deep learning

Deep Learning (DL), a specific form of ML, has recently gained traction by training complex models in large datasets using NNs with several hidden layers. Section 2.5 provides an overview of common DL architectures used for text classification. For applications of DL in the biomedical domain, see some recent surveys written by Wu et al. [12], Hahn and Oleynik [19], Xiao et al. [20], and Shickel et al. [21].

2.2 TRANSFER LEARNING

While DL has promoted a revolution in language modeling in the general domain, it has not yet seen the same success in small data scenarios like the clinical domain. When directly training such complex models in small datasets, they tend to overfit to the dataset and thus not generalize well to new inputs. In order to solve this conundrum, researchers have explored so-called *transfer learning* methods,

in which a model is learned on a larger, typically unlabeled, dataset and then applied to a downstream task, at least partially labeled for evaluation purposes. For instance, language models can be trained on a large collection like Wikipedia to learn the overall structure of human language and then reused on a small private dataset in, e. g., a classification task.

Pan and Yang [22] surveyed transfer learning methods and classified them into four main groups depending on the availability of labels in the source and target domains: (1) when both collections are annotated, one can perform *multi-task learning* to jointly learn a representation that generalizes for both sets; (2) in the specific case of unavailability of labeled data in the target domain, *domain adaptation* can be used to adapt the general domain model to the narrow collection; (3) conversely, if labeled data are available only in the target field, *self-taught learning* [23] can be employed to improve the unsupervised model with annotated data from the target set; (4) finally, if annotated data are not available at all, a good representation scheme can still be learned and transferred to the target collection — also called *unsupervised pre-training* [24], this is the common case of word embeddings [25, Section 15.1] (see Section 2.3.2).

Multi-task learning

Domain adaptation

Self-taught learning

Unsupervised pre-training

Self-training word embeddings may also be attempted if the target dataset is dense enough. The generated model may then be evaluated using either *intrinsic* or *extrinsic* evaluation [26, 27]. In the former method, embeddings are assessed for similarity or relatedness using a reference standard such as UMNSRS [28] or MayoSRS [29]. In the latter case, embeddings are judged by their effect on some downstream task or a reference set of tasks such as the BLUE benchmark [30].

Intrinsic evaluation

Extrinsic evaluation

Since chosen hyperparameters directly affect model quality [31, 32], one should pay attention to reasonable defaults or fine-tune them in the target dataset. For instance, when using word2vec (see Section 2.3.2), the *skip-gram* variant is known to be more appropriate to small datasets and rarer words, while Continuous Bag of Words (CBOW) adapts better to frequent words. Moreover, Chiu et al. [31] showed that: (1) a higher dimensionality (e. g., between 100 and 300 dimensions) requires more data but tends to produce better models; (2) extrinsic tasks benefit from a smaller context window (e. g., 5 words), while larger windows (e. g., 30 words) benefit intrinsic evaluation tasks (since they focus on the topic similarity among words to the detriment of word function); (3) although a larger number of passes over training data (epochs) is beneficial, it tends towards overfitting; (4) a high learning rate may lead to models that do not converge (but are faster to train), while lower learning rates require a longer training time but lead to better models (0.05 seems to be a reasonable default value). Finally, *dropout* should be employed to obtain an average model by randomly dropping neurons at each training step [33].

Dropout

2.3 NATURAL LANGUAGE PROCESSING

When working with textual data, some ML approaches may require further processing steps to first reshape input data (e. g., by simplifying language and thus reducing dimensionality) and then represent words in a shared vector space. In this section, I present common techniques used to approach that.

2.3.1 *Preprocessing Techniques*

A common first step is to remove artifacts out of the text, typically added during typing or introduced by the system as markup language [13, Section 4.2]. Standard cleansing techniques include removal of extra whitespace characters and punctuation markers, while more strict approaches may remove any non-alphanumeric character. Clinical documents may also contain artificial line breaks automatically introduced after a fixed number of characters for visualization, which can be removed by leveraging rules or statistical-based algorithms (e. g., by exploiting relative frequencies of character n-grams).

Text cleansing

After cleaning, one may want to identify sentence boundaries so that contextual information does not span over contiguous sentences. While 90% of full stops (.) determine the end of a sentence in the general domain in English [34], abbreviations (extensively used in the clinical domain, especially in German) may introduce additional noise. Rules or statistical-based approaches are commonly used to address this issue [35].

Sentence detection

Sentences can then be broken down into tokens by using *tokenization* techniques. While it may be intuitive to break sentences into words using whitespaces in Western languages [36], it is not clear how hyphens, punctuation signs, and ambiguous full stops (representing both an abbreviation and a period mark) should be treated [13, Section 4.2.2]. A general approach used is defined by the Unicode Text Segmentation Algorithm[1], implemented, e. g., by the LUCENE[2] programming library.

Tokenization

A simple process that also reduces dimensionality is putting words into lowercase, given that words that appear in the beginning or the middle of sentences usually carry the same meaning. This is, however, a lossy process that may lead to ambiguous interpretation and thus prevent one to differ, e. g., between the two meanings of the word "brown" in *Richard Brown* and *brown paint* [13, Section 4.2.1]. Nevertheless, accepting different cases requires a larger dataset to cover the additional variability.

Lowercasing

Numbers may also pose an issue by artificially inflating the vocabulary size. When they are not necessary, a common strategy is to convert any digit to a fixed digit, e. g., "9". This brings the interesting

1 https://unicode.org/reports/tr29/
2 https://lucene.apache.org

property that numbers still preserve their order of magnitude (e. g., when 12 and 123 are converted to 99 and 999, respectively).

Stopword

Stopwords may also be removed from the input to reduce vocabulary size and further improve performance. Common words typically include prepositions (e. g., "for" and "to"), determiners (e. g., "the" and "a"), and conjunctions (e. g., "and" and "or") — terms that normally do not confer meaning to a sentence. In English, many programming libraries (e. g., LUCENE) internally use the SMART [37] system's list of 524 common words.

2.3.2 *Text Representation*

After preprocessing, text and its constituent words need to be represented in a vector space for downstream tasks. A traditional approach is to use a vector with its size equal to the language vocabulary that is then populated with elements that represent either (a) binary presence of a word (*one-hot encoding*), the term frequency (tf), or the term frequency - inverse document frequency (tf-idf) (the frequency of a word weighted by its rarity in the collection). While in this model words are always orthogonal to each other, more recent developments have led to representations where semantically close words appear nearby, a useful property when dealing with text.

One-hot encoding

Word embeddings

A very popular approach now leverages so-called *word embeddings* (see also Section 2.2) to leverage the contextual information where words occur and represent them in a compressed vector space with typically 200 or 300 dimensions. Word embeddings also allow unsupervised pre-training, when a model is trained on a larger collection and reused in a downstream task (see Section 2.2). Typically, a language model is trained on a large corpus by using NNs and then the hidden layer (the embeddings) are extracted so that words have a stable representation in another problem. For applications of word embeddings to the clinical field, see the surveys written by Wu et al. [12], Kalyan and Sangeetha [26], and Khattak et al. [27].

Word2vec

First proposed by Mikolov et al. [38], WORD2VEC allowed efficient training of two variants of word embeddings, viz. CBOW and *skip-gram*, which predicts a word given its context and vice versa. WORD2VEC's success goes back to negative sampling [39], a method to update only a subset of weights, namely the ones that contribute to a negative prediction. Some time later, Pennington et al. [40] improved over WORD2VEC by also leveraging corpus statistics and released GLOVE. More recently, resilience to out-of-vocabulary words (a common problem in the clinical domain due to misspellings) was added to these models by leveraging subword information, either character n-grams in FASTTEXT [41] or single characters in FLAIR [42].

FastText

Off-the-shelf models pre-trained on large collections have become popular over the years. Pyysalo et al. [43] distributes WORD2VEC embed-

dings trained on PubMed and PubMed Central (PMC) (see Section 3.1.1 for an overview on these datasets), hereafter called BIO.NLPLAB.ORG embeddings; Zhang et al. [44] trained FASTTEXT on MeSH and PubMed and released BIOWORDVEC; later, Chen et al. [45] trained FASTTEXT on PubMed and MIMIC-III and released BIOSENTVEC together with a new version of BIOWORDVEC; FLAIR embeddings were also trained on PubMed. Section 3.1.2 provides a more detailed comparison of the embeddings used in this thesis.

BioWordVec

Embeddings trained on biomedical data seem to achieve better results in this domain: Wang et al. [46] showed that embeddings trained on private clinical notes outperformed others trained on PMC, WIKIPEDIA, and GOOGLE NEWS both quantitatively and qualitatively; similarly, Si et al. [47] demonstrated that WORD2VEC, GLOVE, and FASTTEXT models trained on MIMIC-III improved results on four concept extraction tasks when compared to the same models trained on the general domain. Nonetheless, there seems to be a trade-off between corpus size and content similarity to be evaluated when deciding for such pre-trained models [48]. Note that I explore this specific research question in Chapter 5.

2.4 WORD SENSE DISAMBIGUATION

Expanding short forms (see Chapter 4), sometimes a text preprocessing step, can be seen as a particular form of Word Sense Disambiguation (WSD) (for a survey, see Navigli [49]), a problem defined by Manning and Schütze [13, Chapter 7] as:

> A word is assumed to have a finite number of discrete senses, often given by a dictionary, thesaurus, or other reference sources, and the task of the program is to make a forced choice between these senses for the meaning of each usage of an ambiguous word, based on the context of use.

Techniques proposed to address this problem normally fit into either dictionary-based, supervised, or unsupervised approaches.

In dictionary-based approaches, sense definitions from a dictionary or semantic categories from a thesaurus [50] are typically used. Such approaches leverage the two hypotheses posed by Yarowsky [51] — *one sense per discourse* and *one sense per collocation* — stating that the sense of a given word is mostly constant within a document and that a given sense mainly shares similar phrasal units. A common strategy is then to apply the so-called Yarowsky Algorithm [51], which iteratively uses words from collocations found in training data to build a decision list for sense disambiguation.

In supervised WSD, a dataset annotated with senses is used to train an algorithm, normally framing it as a text classification problem

(see Section 2.5). Here, two strategies may be employed, either (a) leveraging the full context as input to a classification algorithm like Naive Bayes (NB) [52] or (b) picking up the best informative feature from the context using elements from Information Theory [53].

Sense discrimination

Finally, in unsupervised WSD, word contexts are clustered together in a fixed number of groups that could be easily identified by a human, a technique called *sense discrimination*. Unsupervised WSD also allows more fine-grained distinctions that may be required in a narrower domain (such as in medicine), not always with available lexical resources.

Independent of the approach used, context plays an important role. Choueka and Lusignan [54] demonstrated that humans need only a few context words to discriminate among different senses. Moreover, Gale et al. [52] showed that useful information for disambiguation is still found in the context of 50 words on each side and reminiscent information can be found in a context window up to thousands of words. Nonetheless, the agreement among human raters may be highly dependent on skewed distributions of senses typically found in ambiguous words [55, 56], what sets majority vote as a strong baseline [57, 58].

In Chapter 4, I compare unsupervised approaches based on a sense inventory, n-gram, and word embeddings for acronym expansion, a specific case of WSD.

2.5 TEXT CLASSIFICATION

Supervised machine learning techniques can be employed to automatically classify text. According to Manning et al. [59, Section 13.1], a classifier is a function $\gamma : \mathbb{X} \rightarrow \mathbb{C}$ that maps documents in a multidimensional space $\mathbb{X}$ to a set of classes $\mathbb{C}$. The machine learning approach is therefore simply a function $\Gamma(\mathbb{D}) = \gamma$, which runs through a training set of annotated documents $\mathbb{D}$ to achieve a classifier γ. Furthermore, the classifier is known as *one-of*, as it maps only one class for each document.

Shallow learning

Traditional supervised learning approaches (sometimes denominated *shallow learning*) include NB [60], Support Vector Machines (SVMs) [61] (see also Section 5.3.3), and Logistic Regression (LR) [62] (see also Section 5.3.4). Even with the advent of word embeddings, recent work has shown that large-scale text classification using LR provides good results [63].

Convolutional neural network

Regarding DL approaches, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are among the most common architectures used. CNNs leverage convolution operations on a grid-like representation of input data so that nearby cells contribute to a given output. On sequential data like text, Recurrent Neural Networks (RNNs) (such as LSTM) may be employed so that a given state

affects future states (the output information of the last state typically being used as output in classification problems). Gated RNNs expand this concept by allowing the network to forget states accumulated over the sequence. LSTM networks, a popular variant both used for sequence modelling and classification, internalize such memory in a self-loop, whose weight is also learned during training (for a more detailed description on LSTM, see Section 5.3.5).

Long short-term memory

Alternatively, the creation and maintenance of hand-crafted rules is mostly labor-intensive, but can lead to very good results that surpass automated mechanisms [64]. They are interesting when (a) the dataset is small, (b) domain specialists with skills in logic and regular expressions are available, and/or (c) explainability is desired or required. However, the effort required to create and maintain such a set of rules, generally based on regular expressions, is not negligible [59, Section 15.3]. Moreover, the content of each class may change with time [59, Section 13.4], which requires a constant maintenance effort.

In Chapter 5, I explore DL methods together with traditional approaches such as LR, SVM, and a Rule-Based Classifier (RBC) for clinical text classification.

2.6 INFORMATION RETRIEVAL

Information Retrieval (IR) is the process of obtaining resources from a large collection of textual documents that match an *information need* [59, Section 1.1]. Computer systems designed to aid in this process by reducing information overload are called *search engines*, where an information need is translated by the user into a textual query, normally a sequence of keywords. A combination of scientific progress, high-level engineering, declining costs of computer hardware, and commercial applications of search engines (by the usage of tailored advertising) has allowed the field to progress even in the face of exponentially growing collections like the web [59, Preface].

From the engineering point of view, document representation via a structure known as *inverted index* avoids the need of scanning all documents in a collection for a given query. It is based on an ordered map that associates words from the collection vocabulary to the documents where they occur. As a result, searches can be performed in logarithmic time with respect to the vocabulary size, which is typically smaller than the collection size.

Inverted index

Lastly, ranking functions such as tf-idf allow the ordering of search results by relevance, which is essential for shielding the end-user from being swamped with contents of limited relevance. The basic idea comes from the sense that documents with higher tf of the query expression are more relevant to the query. Moreover, words or phrases contained in many documents (such as "disease" in a medical corpus) should have a penalty proportional to the number of documents

Tf-idf

with the word in the collection, its reciprocal being named inverse document frequency (idf).

Precision-recall tradeoff

An IR system has to find the right balance between precision and recall; this is also called the *precision-recall tradeoff* (see Chapter 6 for a method proposed by us to strike the right balance). Improving recall is normally done via *query expansion*, a process by which a given user query is expanded with related words obtained either from the search results themselves (local) or external sources (global) [59, Chapter 9].

Query expansion

Global methods for query expansion involve either (a) cleaning query/document text [59, Chapter 3] or (b) the usage of thesauri/terminologies [59, Section 9.2]. With text preprocessing techniques (see Section 2.3.1), one can obtain more documents that would not have been retrieved by the original query. Conversely, terminologies can be leveraged to obtain synonyms, hypernyms, or meronyms[3]. Terminologies can be either manually curated (e. g., SNOMED CT) or automatically created by leveraging *distributional semantics* (e. g., using word co-occurrence statistics). Optionally, dimensionality reduction techniques like Latent Semantic Analysis (LSA) or word embeddings can also be employed for query expansion by providing similar terms to an input query.

These methods mostly improve recall, but in some cases also precision. However, precision can easily be improved by using so-called boosting terms that either increase (or decrease) the score of documents containing chosen (non-)relevant terms. Precision is also improved by the usage of Learning to Rank (LTR) algorithms that apply (mostly) supervised ML strategies to re-rank results based on a training dataset [65].

In Chapter 6, I explore query expansion based on terminologies, handcrafted rules, and word embeddings for an IR scenario in precision medicine.

3 A semantic relationship describing a constituent part of or a member of something.

3 MATERIALS AND METHODS

3.1 MATERIALS

Since clinical data is normally not publicly shared due to ethical concerns, I employed transfer learning methods (see Section 2.2) to overcome this limitation. In this section, I present the public datasets I used throughout the thesis and in which word embeddings were pre-trained. Further data used for specific experiments is described at the beginning of the corresponding chapter.

3.1.1 *Datasets*

MEDLINE

A commonly employed resource is the MEDLINE[1] collection maintained by the National Library of Medicine (NLM). MEDLINE is a database of millions of bibliographic records from the fields of life sciences and bio-medicine, the majority including the publication abstract and enriched with manually assigned Medical Subject Headings (MeSH) terms [66] (hereafter denominated "document"). PubMed, maintained by the NLM and the National Institutes of Health (NIH), provides online access to both MEDLINE documents and unchecked non-MEDLINE content as supplied by the publisher. A baseline file[2] released yearly contains both MEDLINE and quality-checked non-MEDLINE documents and is the primary source of data.

PubMed

PMC

PubMed is also commonly associated with PubMed Central (PMC) open-access data. PMC is a digital repository of full-text articles in the same fields as PubMed and maintained by the same institutions since 2000, the vast majority also indexed for PubMed. While all papers published in PMC are available in their entirety, only a subset of them is released under a real open access license. This open-access subset is the de-facto collection used for research [67].

MIMIC-III

Complementary to biomedical literature, de-identified clinical texts are available in the MIMIC-III corpus [68]. MIMIC-III contains data from 61 532 intensive care unit stays of 46 520 patients from the Beth Israel Deaconess Medical Center in the United States. Textual notes (discharge summaries, electrocardiogram reports, imaging reports, and nursing notes) are available in the `NOTEEVENTS` table, the common training material used for downstream Natural Language Processing (NLP) tasks.

1 https://www.nlm.nih.gov/bsd/medline.html
2 https://www.nlm.nih.gov/bsd/licensee/baseline.html

Table 3.1 provides a quick comparison of these datasets regarding size in the number of documents, sentences, and tokens. For a detailed overview of digital libraries, see Hersh [66].

	PUBMED	PMC	MIMIC-III
Documents	$\sim 25.4 \times 10^6$	673×10^3	2.08×10^6
Sentences	$\sim 153 \times 10^6$	105×10^6	41.7×10^6
Tokens	$\sim 3.63 \times 10^9$	2.59×10^9	539×10^6

Table 3.1: Overview of public datasets used in this thesis.

3.1.2 *Pre-trained Word Embeddings*

Instead of retraining models from scratch, I leveraged word embeddings (see Section 2.3.2) that had been pre-trained in the collections described in the previous section. Table 3.2 compares the two chosen models, namely BIOWORDVEC [45] and BIO.NLPLAB.ORG [43].

	BIOWORDVEC	BIO.NLPLAB.ORG
Base model	FASTTEXT	WORD2VEC (skip-gram)
Training corpus	PUBMED + MIMIC-III	PUBMED + PMC
Dimensions	200	200

Table 3.2: Pre-trained word embeddings used in this thesis.

While both models leverage PUBMED as the main source of texts and employ the same dimensionality (200) for the released embeddings, they differ in their intended goals. Overall, BIOWORDVEC is more suitable for the clinical domain due to its use of FASTTEXT — a model including subword information and thus resilient to misspellings — and the association of clinical data from MIMIC-III. Conversely, BIO.NLPLAB.ORG embeddings leverage the full text of documents from PMC to enrich PUBMED data and — since peer-reviewed scientific texts are less prone to misspellings — employ the traditional WORD2VEC tool to produce embeddings useful to the broader biomedical domain.

3.2 EVALUATION METRICS

Studies using Machine Learning (ML) must be consistently evaluated to ensure result comparability, a task commonly done using standard quality measures such as accuracy, precision, recall, and F-score (for a review of metrics, see Hand [69]). Example data is commonly divided into two sets, one for training and a second one for testing. The split is normally done in a way that the training set contains 80% to 90% of the available data and the part the algorithm uses for learning, while the test set is used for evaluation.

The current section describes the overall evaluation metrics used in this work. For the sake of simplicity, we refer to the documentation unit as "document"; depending on the use case, it may refer either to a single document (e. g., a MEDLINE record containing the text abstract) or a collection of records associated with a given patient.

3.2.1 *Accuracy, Precision and Recall*

Accuracy

A naive approach to measuring the quality of an experiment is the accuracy. Accuracy is simply the ratio between the number of correct documents and the total number of documents, as seen in Equation 3.1.

$$A = \frac{\text{number of correct documents}}{\text{total number of documents}} \tag{3.1}$$

Accuracy can be calculated over documents retrieved for a query in an Information Retrieval (IR) scenario or, for classification problems, over correctly classified documents. However, accuracy can be artificially high if very few documents are retrieved since classes are highly unbalanced. To overcome this problem, results are commonly expressed by precision (also known as positive predictive value) and recall (also called sensitivity). The definition is guided through a contingency table, defined in Table 3.3.

	RELEVANT	NON-RELEVANT
RETRIEVED	true positive (tp)	false positive (fp)
NON-RETRIEVED	false negative (fn)	true negative (tn)

Table 3.3: Contingency table.

Precision
Recall

Based on this contingency table, precision and recall metrics can be easily defined in Equation 3.2 and Equation 3.3. Precision is the relevant fraction of the retrieved results, while recall is the retrieved fraction of the relevant results [13, Section 8.1].

$$P = \frac{tp}{tp + fp} \tag{3.2}$$

$$R = \frac{tp}{tp + fn} \tag{3.3}$$

Accuracy follows then naturally in Equation 3.4.

$$A = \frac{\text{number of correct documents}}{\text{total number of documents}} = \frac{tp + tn}{tp + fp + fn + tn} \tag{3.4}$$

3.2.2 *F-score*

Although precision and recall solve the quality assessment in the IR context, another problem arises: there is no single metric anymore to measure the outcome of experiments in a unique way. For instance, a small drop in precision might be traded for a much higher recall or vice versa. Even though the relative weight between the measures is context-dependent, it is possible to express them together via a metric known as F-score.

F-score

F-score is defined by Manning et al. [59, Section 8.1] as the harmonic mean between precision and recall and is expressed via Equation 3.5.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{3.5}$$

Here, α is the weight factor between precision and recall, related to β via Equation 3.6.

$$\beta^2 = \frac{1-\alpha}{\alpha} \tag{3.6}$$

With equal weight ascribed to precision and recall, F-score is called F_1, considering that $\beta = 1$ ($\alpha = 0.5$). The expression may then be further simplified to Equation 3.7.

$$F_{\beta=1} = \frac{2PR}{P+R} \tag{3.7}$$

3.3 METHODS

3.3.1 *Superior and Inferior Limits*

Minimum and maximum algorithm efficiency should also be estimated to set sensible limits for the quality of results, thus marking the interval outside of which results are not expected. The inferior limit (also known as baseline) normally refers to the outcome of the simplest strategy to solve a problem without too much effort [13, Section 7.1.3], usually a majority approach or an established method used as control.

Baseline

The superior limit of an algorithm efficiency is related to the fact that even human specialists never completely agree on correct labeling or class assignment, due to lack of information, fuzzy criteria, context-dependency, or even by mistake. It is thus a reasonable assumption that a computer strategy trained on human-generated data cannot perform better than the humans who created the data. Disagreement rates between two raters can be determined by Cohen's *kappa* (κ) index, which rates how much an agreement between annotators is higher than an agreement by chance.

Kappa index

Cohen's κ is mathematically defined via Equation 3.8, where $P(A)$ is the number of times the experts agreed and $P(E)$, the number of times it was expected they would agree by chance [59, Section 8.5]. The index approaches 1.0 the more they agree, 0.0 if they just agree by chance or a negative number if they are worse than random selection. Manning et al. [59, Section 8.5] empirically reported that values higher than 0.8 are good, between 0.67 and 0.8 are fair, and lower than 0.67 are dubious.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.8}$$

3.3.2 *Statistical Significance Testing*

When comparing the results of new methods to a baseline, statistical hypothesis testing should be performed to ensure any found difference is not due to chance. Since runs in a ML experiment typically use the same dataset, observed results are not independent and thus a paired test is normally required.

In cases where the variable is qualitative (e. g., correct counts in Chapter 4 and Chapter 5), McNemar's test is recommended given the low number of type I errors [70]; for quantitative variables that can be assumed to have a normal distribution, a t-test for paired samples is commonly recommended; for non-parametric quantitative variables (e. g., recall in Chapter 6), Fisher's randomization test has been favored over the traditional Wilcoxon signed-rank test given the availability of computing resources [71].

McNemar's test

Fisher's randomization test

Throughout this thesis, I set the null hypothesis that there is no difference between a given experiment and the baseline (two-tailed), with the significance level of $\alpha = 0.05$. I then run significance tests to determine the probability, under the null hypothesis, of obtaining results at least as extreme as the one observed (p-value) and call results statistically significant when $p \leqslant \alpha$.

4

CLINICAL TEXT CLEANSING

In this chapter we explore unsupervised methods based on n-grams and word embeddings to expand short forms such as acronyms, an important step in clinical text cleansing to improve text legibility and allow further automated processing.

Using data from the Austrian public healthcare provider KAGes, we propose a minimal set of filtering rules that improve precision of acronym expansion and I show that its association with word embeddings outperforms traditional approaches based on n-grams and sense inventories. A preliminary version of this work applied to abbreviations was published at MEDINFO 2017.

This chapter had contributions from Johannes Hellrich and Stefan Schulz. The acronym detection algorithm, the filtering rules, and the n-gram method were first proposed by SS. SS also contributed to annotating the expansion standard. The idea of using word embeddings for acronym expansion was first discussed with JH when I was at the JULIE Lab (Jena, Germany) in June 2018.

4.1 INTRODUCTION

4.1.1 *Motivation*

One of the characteristics of clinical texts is the widely use of short forms such as abbreviations and acronyms [72]. While in scientific publications the introduction of a short form together with its expansion using so-called acronym-definition pairs (e. g., the ones used in this thesis) is considered a good practice [73], short forms in clinical texts tend to be seldom defined, as their meaning is supposed to be known by the reader (e. g., *AIDS* → *Acquired Immune Deficiency Syndrome*).

However, even though short forms can normally be understood by staff from the same clinical department, the same does not always happen across clinical specialties, professional groups (e. g., doctors and nurses), institutions, and clinicians with different mother tongues. In a study with pediatric notes, Sheppard et al. [74] showed that while pediatric doctors could understand 56–94% of the short forms present therein, other healthcare professionals could recognize only 32–63%.

Expanding such short forms would furthermore improve legibility for patients, parents, and caregivers, sometimes in charge of conveying key information across healthcare providers [75]. In this process, acronyms were shown to be the major cause of misunderstanding for laypeople, independent of health literacy [76]. Indeed, the Joint

Commission International, a non-profit organization that accredits hospitals in the United States and thus allows them to obtain public reimbursements, recommends in their standards[1] that:

> Abbreviations are not used on informed consent and patient rights documents, discharge instructions, discharge summaries, and other documents patients and families receive from the hospital about the patient's care.

Term conflation

Overall, short form resolution is also important for *term conflation* [77], when several terms that share a synonymic relation are mapped to a single concept, e. g., *{ECG, EKG, Electrocardiogram, Elektrokardiogramm} → Electrocardiogram*. Term conflation aids automated systems that are still syntax-based — such as out-of-the-box search engines (see Chapter 6), concept mappers, and text classifiers (see Chapter 5) — by reducing the input dimensionality.

Abbreviation

Acronym

Short forms happen in a myriad of ways. *Abbreviations* are shortened versions of words, marked with a period mark in many languages (e. g., *fig.* → *figure*). *Acronyms* are a special kind of abbreviations composed of the initial characters of a phrase or syllables/morphemes from a word. Acronyms can further be subclassified into single word (e. g., *ECG* → *Electrocardiogram*) and multiword acronyms (e. g., *DM* → *Diabetes Mellitus*). Especially in languages that allow single-word composition like German, acronyms are quite commonly used to shorten long words (e. g., *PAE* → *Pulmonalarterienembolie*, pulmonary embolism) and, in this case, each acronym character maps to one morpheme in the expansion.

Sense inventory

Clinical sense inventories mapping acronyms to their senses are widely available in English, with two resources in the Unified Medical Language System (UMLS) alone: the LRABR table and the Specialist Lexicon. Such resources cover approximately two-thirds of short forms found in biomedical texts [78, 79]. They can be further expanded using online databases of scientific literature such as PubMed, giving origin to automatically maintainable databases such as ADAM [79] and ALICE [80].

Detection

Expansion

Disambiguation

Resolving acronyms is traditionally subdivided into three sub-tasks, namely detection, expansion, and disambiguation. Acronym detection is the task of deciding whether a given sequence of characters is an acronym or not, normally by looking at the casing, presence of a nearby expansion, or presence in a sense inventory. Expansion refers to mapping a given acronym to candidate forms, found either in a sense inventory or nearby in the source text (e. g., in an acronym-definition pair). Finally, disambiguation refers to the task of finding the proper expansion of an acronym given its lateral context (i. e., the words appearing right and left to the word).

1 https://www.jointcommissioninternational.org/assets/3/7/JCI_Standards_Interpretation_FAQs.pdf

Most strategies for acronym expansion are based on the distributional hypothesis, according to which words with similar meaning occur in similar contexts[2]. Therefore, this context can be used to predict the meaning of a given word or, conversely, to map a short form to its expansion. *Predictive models* such as WORD2VEC (see Section 2.3.2) tend to improve upon *count-based models* (such as Latent Semantic Analysis (LSA) or raw co-occurrence counting) because they use self-supervision to optimally determine the context a given word occurs, a process that also makes the model denser and thus better/faster [82]. While acronym disambiguation can be seen as a particular form of Word Sense Disambiguation (WSD) (see Section 2.4), I adapted these techniques for *acronym expansion*, the focus of this thesis chapter.

Distributional hypothesis

Predictive model

Count-based model

4.1.2 *Related Work*

Table 4.1 shows a chronological overview of acronym resolution approaches in both the broader biomedical domain (e. g., literature articles from PUBMED) and the more specific clinical domain (patient data). Note that even though clinical texts can be highly distinct from biomedical literature articles, they still share a similar vocabulary when compared to, e. g., patent/legal texts. For each work, I detailed the method used, the usage of a sense inventory and the reported efficiency. For a more comprehensive review on the topic, see also Spasic [77].

AUTHOR	YEAR	DOMAIN	LANGUAGE	METHOD	SI	F_1 / A
Liu et al. [83]	2001	Clinical	English	Hybrid	Yes	0.97†
Pakhomov [84]	2002	Clinical	English	Hybrid	Yes	0.89†
Wu et al. [85]	2013	Clinical	English	Hybrid	Yes	0.72†
Siklósi et al. [86]	2014	Clinical	Hungarian	Unsupervised	Yes	0.85†
Mowery et al. [58]	2016	Clinical	English	Unsupervised	Yes	0.69†
Kirchhoff et al. [87]	2016	Clinical	English	Unsupervised	Yes	0.72†
Charbonnier et al. [88]	2018	Biomedical	English	Unsupervised	No	0.86†
León [89]	2018	Biomedical	Spanish	Unsupervised	No	0.83§
Spasic [77]	2018	Biomedical	English	Unsupervised	No	N/A

Table 4.1: Comparison with related work on acronym resolution. SI = Sense Inventory. † denotes accuracy and § denotes F_1-score.

Liu et al. [83] explored a hybrid approach consisting of two phases: (1) unsupervised string matching using text from UMLS concepts; and (2) supervised classification using either Naive Bayes (NB), decision list, or k-Nearest Neighbor (kNN) with nearby words as features. With medical reports from the New York Presbyterian Hospital, they reported experiments with different window sizes and classification

2 This was popularized by the Firth [81] quote "You shall know a word by the company it keeps".

algorithms and obtained a global optimum of 0.97 accuracy with NB and ten nearby stemmed words.

Pakhomov [84] employed a similar, semi-supervised, method on six abbreviations commonly found in rheumatology notes from the Mayo Clinic. First, a training corpus was automatically generated by identifying expansions from the LRABR table in the collection and mapping them to their acronyms. The lateral context on which the expansion was found was then fed together with the acronym to train a Maximum Entropy (MaxEnt) model. The author reported accuracy of 0.89 for the best model created.

The 2013 ShARe/CLEF eHealth Evaluation Lab [57] held a track on normalization of short forms to UMLS concepts. The challenge used texts from the MIMIC-II dataset, annotated with text spans and UMLS codes. The winning system [85] obtained an accuracy of 0.72 with a hybrid system consisting of three steps: (1) construction of the sense inventory out of the training data and terminologies such as LRABR; (2) disambiguation by either Support Vector Machine (SVM) classifiers, a profile-based method leveraging a vector space model, or a majority-based approach; and (3) UMLS encoding using the data from the training dataset.

Siklósi et al. [86] explored the impact of an unsupervised approach. They experimented with both an external and an internal lexicon to increase the efficiency of short form resolution on Hungarian ophthalmological notes. Mapping was performed using regular expressions matched against the lexicons and the corpus itself, with a calculated F_1 of 0.85.

Mowery et al. [58] analyzed 2013 CLEF participants' systems in detail and concluded that a majority approach would perform second best with 0.69 accuracy, since 81% of short forms were unambiguous and out of the forms with a low ambiguity, one of the meanings has an average incidence of over 80% alone.

Kirchhoff and Turner [87] experimented with web mining to obtain expansion lists and then used them to disambiguate short forms found in nurse notes. A first step collected expansion tables found on specialized websites, like WIKIPEDIA and the National Institutes of Health (NIH). A second step mapped short forms to different expansions and combinations thereof using a phrase-based statistical machine translation system. A third step then scored candidates using a document-level language model based on recurrent neural networks. They reported a final F_1 of 0.72.

Charbonnier and Wartena [88] used word embeddings to disambiguate acronyms from image captions of scientific papers. They trained embeddings on all words from the corpus and compared the mean context vector of words in the expansion with the vectors of words in the context of an acronym. They reported that this approach outperforms classical cosine similarity and reaches an accuracy of 0.86.

They also showed that word embeddings trained on the same domain had better results than embeddings trained on larger corpora in the general domain.

The 2018 Biomedical Abbreviation Recognition and Resolution (BARR) shared task [90] promoted the development of systems to identify acronym-definition pairs (Task 1) and expand short forms to their full form (Task 2). BARR used textual data from Spanish health records. The winning system [89] proposed for Task 2 two unsupervised approaches, namely one based on rules stated in a context-free grammar and other based on n-gram lists ($1 \leqslant n \leqslant 3$) (bag-of-words), with which they achieved F_1 of 0.83.

Finally, Spasic [77] employed lexico-syntactic filtering to map acronyms to candidate expansions extracted using patterns of part-of-speech tags. She reports a 32 relative recall increase in an information retrieval scenario consisting of five different subsets of biomedical corpora, such as PubMed and 12B2, but did not report overall accuracy or F_1.

4.1.3 *Research Problem*

Acronyms in clinical texts are widely used but seldom accompanied by their expansion, which not always can be found in sense inventories — especially in languages with fewer lexical resources. However, automatically expanding such short forms is important to health professionals, patients, and systems that process clinical text for secondary uses.

A great deal of research has been conducted on the acronym expansion problem, using different strategies applied to several languages and types of data, but few experiments have been carried out without an annotated dataset and even fewer without a preliminary sense inventory. Moreover, all unsupervised studies not relying on a sense inventory were performed in the broader biomedical domain, where acronym-definition pairs are commonly found.

Therefore, I propose here an unsupervised strategy to address the acronym expansion problem in German without a preexisting sense inventory. I hypothesize that a predictive model like word embeddings outperforms a count-based model like n-grams. To the best of my knowledge, this work is the first to apply an unsupervised technique in the clinical domain without relying on a sense inventory.

4.2 MATERIALS

4.2.1 *Dataset*

We obtained a collection of 30 000 de-identified discharge summaries written in German from the department of cardiology of KAGes, the public healthcare provider of the province of Styria, Austria. The

documents had been produced during clinical routine from 2009 to 2014.

We split the available data into 90% for training and 10% for test (see Table 4.2). I trained models and fine-tuned hyper-parameters using an evaluation dataset built out of the training set and I report results using a subset of 104 acronyms randomly selected from the test set.

	TRAINING	TEST
Characters	58 444 275	5 950 649
Types	270 081	60 580
Tokens	8 600 873	879 993
Acronyms	679 566	74 229
Unique acronyms	11 711	3109
Sentences	2 546 129	224 967
Documents	28 729	3192

Table 4.2: Description of the dataset used for acronym expansion experiments.

4.2.2 *Detection Standard*

Even though acronym detection was not the focus of this work, we created a stricter version of the algorithm proposed by Park and Byrd [91] so that we could automatically build an evaluation standard, as well as filter out expansions containing other acronyms (see Section 4.3.2). We thus consider a candidate acronym if, and only if, it contains (1) from 2 to 7 alphanumeric characters, and (2) more upper case than lower case characters. Additionally, I manually assessed the validity of all sampled acronym types before expansion.

4.2.3 *Expansion Standard*

Of the remaining valid acronyms, we manually assessed the top-10 expansions given by the proposed methods using the scale I proposed below, whose usage is bound by the annotation guidelines depicted in Table 4.3.

- 2 (full match): the expansion relates to the acronym and fully expands it;
- 1 (partial match): the expansion relates to the acronym, but does not fully expand it and thus needs further processing;
- 0 (no match): the expansion does not relate to the acronym.

EXPANSION	EXAMPLE	ASSESSMENT
Perfect match	*SR* → *Sinusrhythmus*	2
Domain character change	*TX* → *Transplantat*	2
Hyphenation	*CK* → *Creatinin-Kinase*	2
Plural	*DES* → *Drug Eluting Stents*	2
Grammatical case	*HA* → *Hausarztes*	2
Spelling variant	*CV* → *Kardioversion*	2
Capitalization	*DES* → *Drug eluting Stent*	2
Synonym	*ASS* → *Aspirin*	1
Semantically equivalent translation	*EF* → *Auswurffraktion*	1
Abbreviation	*LV* → *Linksventr*	1
Spelling error	*CA* → *Cornarangiographie*	1
Longer (without change in meaning)	*BZ* → *Blutzuckerwerten*	1
Not a match	*PCI* → *Stenting*	0
Hypernym	*VVI* → *Herzschrittmachermodus*	0
Longer (with change in meaning)	*HA* → *Hauszahnarzt*	0

Table 4.3: Annotation guidelines used when creating the acronym expansion standard.

4.2.4 *Evaluation Metrics*

I evaluated the proposed strategies using precision (P) and recall (R) following the SENSEVAL[3] evaluation tasks. Precision is defined as the ratio of correct expansions to the generated expansions, while recall is the ratio of correct expansions to all acronyms [92].

$$P = \frac{\#\text{correct}}{\#\text{found}} \tag{4.1}$$

$$R = \frac{\#\text{correct}}{\#\text{all}} \tag{4.2}$$

Note that in the SENSEVAL definition, an algorithm can choose not to expand a given instance and thus #found $\leqslant$ #all, which implies that $R \leqslant P$. Moreover, an algorithm can provide a ranked list of possible expansions, of which more than one may be correct. I report precision, recall, and F_1 at rank $k = 1$ to evaluate the scenario where an acronym is automatically mapped to the most probable expansion candidate. I also report results using both a strict matching (expansions assessed with 2) and a lenient matching (expansions assessed with either 2 or 1).

3 http://web.eecs.umich.edu/~mihalcea/senseval/

4.3 METHODS

Using a minimal number of filtering rules, I compared a predictive model based on word embeddings to a count-based model based on n-grams, set the baseline to a majority approach, and also evaluated the coverage offered by a clinical German sense inventory available on WIKIPEDIA. I performed paired McNemar's tests (see Section 3.3.2) comparing correct counts of each method to the baseline to check whether any differences were statistically significant, with $\alpha = 0.05$. Our source code is available at GitHub[4] under version 2 of the Apache License.

4.3.1 *Preprocessing*

To reduce data sparseness, we preprocessed our data using the techniques described in Section 2.3.1, namely removal of non-alphanumeric characters (including line breaks) and digit normalization.

4.3.2 *Filtering Rules*

We employed a minimal set of handcrafted binary rules to reduce the number of possible candidates and increase precision for all strategies described next. Filtering techniques were applied to full forms alone and in combination with the queried acronym. If present, acronyms were trimmed of a trailing 's' (plural form) and 'x' (commonly used to designate a transplant, e. g., *NTx* → *Nierentransplantation*).

For the expansion E itself, we devised the following restrictions:

- **Noun**: it must have at least one noun, marked as an initial uppercase in German (e. g., *ambulanz* is invalid);
- **Acronym**: it must not contain another acronym (e. g., *Thrombo ASS* in *TASS* → *Thrombo ASS* is invalid[5]).

For the acronym-expansion A, E pairs, we applied the following heuristics:

- **Relative length**: an expansion can be up to 20 times longer than the acronym and the acronym has to be at most 60% of the full form length[6] (e. g., *RR* → *Risikofaktorenmanagement insbesondere regelmäßige* is invalid);
- $\mathbf{lev}(\mathbf{A}, \mathbf{A}') < |\mathbf{A}|$: the Levenshtein distance of the acronym A to an acronym A' automatically generated from the expansion[7]

4 https://github.com/bst-mug/acres

5 Note that I consider *TASS* → *Thrombo ASS* a partial expansion, since *ASS* can be further expanded to "Acetylsalicylsäure".

6 Relative length bounds were empirically determined using the training set.

7 The acronym A' is automatically generated from a candidate expansion by concatenating the initials of each word that is not a preposition.

should be less than the acronym length (e. g., *VB* → *Vorgeschichte darf als bekannt* is invalid);

- **Valid expansion**: the expansion must start with the first acronym character and contain all remaining acronym characters in the same order (e. g., *MIN* → *Implantation* is invalid);
- **Last acronym character in the last word**: the last character of the acronym must be in the last word, but not as the last character (e. g., *AIN* → *Ablation* is invalid);
- **The last word is a sub-acronym**: in multi-word expansions, the initial of the last word of the expansion must be in the acronym *and* the remaining acronym characters must still match (e. g., *TIA* → *Transmuraler Myokardinfarkt* is invalid).

4.3.3 *Acronym Expansion*

Baseline

I set the baseline to a majority approach. In this strategy, I expanded each acronym to the most common n-gram in the collection that passed our filtering rules described before.

Using a Sense Inventory

I also compared the proposed methods to a German medical sense inventory extracted from Wikipedia[8] and manually curated by a medical doctor. This allowed me to evaluate the coverage of a crowd-sourced, manually curated, sense inventory for the task of acronym expansion in a narrow clinical domain.

Using N-grams

We experimented with a count-based method which used the frequency of n-grams as a signal for ranking. Given an acronym type, we (1) looked at the training data for lateral contexts of decreasing frequency where the given acronym appeared and then (2) retrieved candidate expansions that appeared within the same found context. Since we noticed in preliminary experiments in the training set that the right context is more important than the left (mostly explained by test values, e. g. *RR 120/80mmHg*), we first retrieved matches using a longer right context.

Using Word Embeddings

We also experimented with a predictive method using word embeddings. Given a vector representation of an acronym, our main strategy

8 https://de.wikipedia.org/wiki/Medizinische_Abkürzungen

comprised of finding the word (or collocation of words) whose vector was closest to the given vector and which passed the filtering rules (see Figure 4.1).

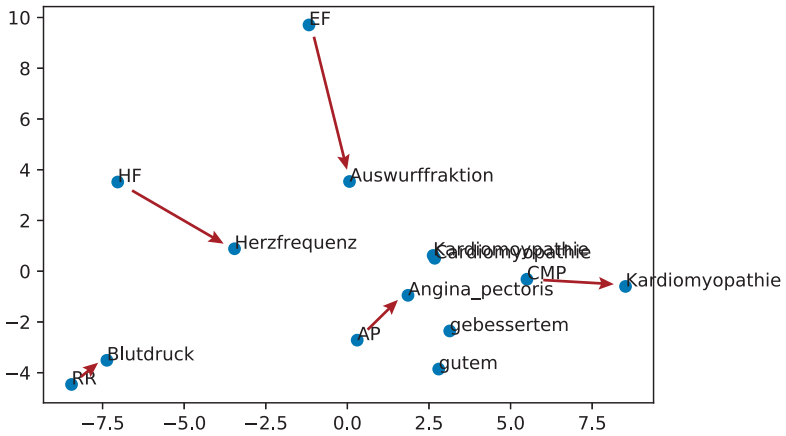

Figure 4.1: 2D Principal Component Analysis (PCA) representation of the model based on word embeddings (example words).

Multiword acronyms span several tokens and thus required special handling. Following Mikolov et al. [39], I captured common collocations of two words[9] and added them to the model, thus allowing, e. g., the expansion *AP* → *Angina Pectoris*. I employed a Continuous Bag of Words (CBOW) model (see Section 2.3.2), which had been known to perform better with small datasets because there are fewer parameters to be learned.

I trained the models using Pyhton's gensim[10] implementation of WORD2VEC and fine-tuned hyperparameters with grid search. The best model had a window size of 2 over the central word and was trained using a neural network with a hidden layer containing 100 neurons, learning rate $\alpha = 0.025$, and negative sampling out of five noise words.

As I did not need to disambiguate acronyms (see Section 4.4.2), I worked at the type level (i. e., the abstract representation of a word, not its use in a specific context) and therefore used only the most similar vector of each acronym as an expansion candidate, further filtered by the rules described before in Section 4.3.2.

4.4 RESULTS

4.4.1 *Detection Standard*

Table 4.4 shows the number and ratio of acronym types in each identified acronym group. Out of an initial set of 211 acronym types selected from both the training and test splits, I considered 140 (66.4%) Ger-

9 See Section 4.5.1 for a discussion on the impact of longer collocations.
10 https://radimrehurek.com/gensim/models/word2vec.html

man acronym types valid for expansion. I did not try to expand (a) lexicalized acronyms[11] in English (e. g., *COPD* → *Chronic obstructive pulmonary disease*) or Latin (e. g., *CX* → *Circumflexus*); (b) a sequence of letters without any specific meaning (e. g., *QRS*[12]); (c) roman numbers; (d) coding schemes (e. g., *VVI*[13]); (e) section titles; or (f) acronyms containing tokenization errors.

ACRONYM GROUP	EXAMPLE	COUNT	RATIO
Valid German	*EKG*	140	66.4%
English (lexicalized)	*COPD*	38	18.0%
No specific meaning	*QRS*	8	3.8%
Latin (lexicalized)	*CX*	7	3.3%
Roman number	*III*	7	3.3%
Coding scheme	*VVI*	6	2.8%
Section title	*BEFUND*	3	1.4%
Tokenization error	*LDH209*	2	0.9%
Total		211	100%

Table 4.4: Counts of acronym types after validity restrictions.

4.4.2 *Expansion Standard*

The final evaluation dataset contains 4666 assessments, of which 385 expansions (8.24%) fully match and 332 (7.11%) partially match the acronym. Cohen's kappa (see Section 3.3.1) between the two annotators is 0.85, which indicates almost perfect agreement. Moreover, acronyms in this collection are rarely ambiguous, with only seven out of 195 (3.59%) with more than one sense:

- *ACC* → *{Acetylcystein, Arteria carotis communis}*
- *AP* → *{Alkalische Phosphatase, Angina pectoris}*
- *AT* → *{Augentropfen, atriale Tachykardie}*
- *DM* → *{Diabetes Mellitus, Durchmesser}*
- *HA* → *{Hausarzt, Herzaktion}*
- *HK* → *{Herzkrankheit, Hyperkinesie}*

11 Even though acronyms in English and Latin can be found in German clinical texts and are accepted as part of the language, they are mostly already present in a sense inventory in the source language and thus do not constitute the focus of the present work.

12 *QRS* corresponds to the three major inflection points typically seen on an electrocardiogram.

13 *VVI* is a code denoting a common pacemaker mode; each letter denotes, in sequence, the heart chamber that is paced ("V" refers to ventricle), sensed, and the response to a sensed event ("I" stands for inhibition).

- *VB* → *{Venenbypass Phosphatase, Vorbefund}*

Table 4.5 presents the results of acronym expansion using strict and lenient matching. With either strict or lenient matching, n-gram is the strategy that provides the largest number of expansions (104). However, since a comparatively low number of expansions is correct (46 and 51, respectively), it produced low precision (and recall) metrics. Conversely, WORD2VEC obtained the highest number of correct expansions with a smaller number of candidate expansions, thus producing the highest metrics using lenient matching and, except for precision, also using strict matching. All strategies produced results significantly above the majority baseline ($p < 0.001$). When compared to the sense inventory, only WORD2VEC with lenient matching had significantly better results ($p < 0.05$); this strategy also had significantly better results ($p < 0.05$) than n-gram matching. Overall, except for #correct and recall, n-gram has most results below the fixed sense inventory ($p > 0.05$).

MATCHING	STRATEGY	#FOUND	#CORRECT	P	R	F_1
Strict	Majority	104	15	0.14	0.14	0.14
	Sense inventory	64	44	0.69	0.42	0.52
	n-gram	104	46	0.44	0.44	0.44
	WORD2VEC	84	54	0.64	0.52	0.57
Lenient	Majority	104	17	0.16	0.16	0.16
	Sense inventory	64	47	0.73	0.45	0.56
	n-gram	104	51	0.49	0.49	0.49
	WORD2VEC	84	63	0.75	0.61	0.67

Table 4.5: Number of found and correct acronyms, precision (P), recall (R), and F_1-score with different strategies at rank $k = 1$ with both strict and lenient matching.

Figure 4.2 depicts F_1 metrics at different ranks k for the four evaluated metrics and both strict and lenient matching. The plot shows that count-based methods like majority and n-gram increasingly benefit from an evaluation at higher ranks, where a correct match is expected by chance. Meanwhile, both the sense inventory and the predict-based method (WORD2VEC) sustain a constant efficiency at all ranks.

4.5 DISCUSSION

4.5.1 *Error Analysis*

In order to better understand the behavior of WORD2VEC, I also qualitatively evaluated the embeddings generated thereby. Table 4.6 shows that the model was able to capture the meaning behind the adjective "gutem" (good), including the proper "em" inflection. Moreover, it

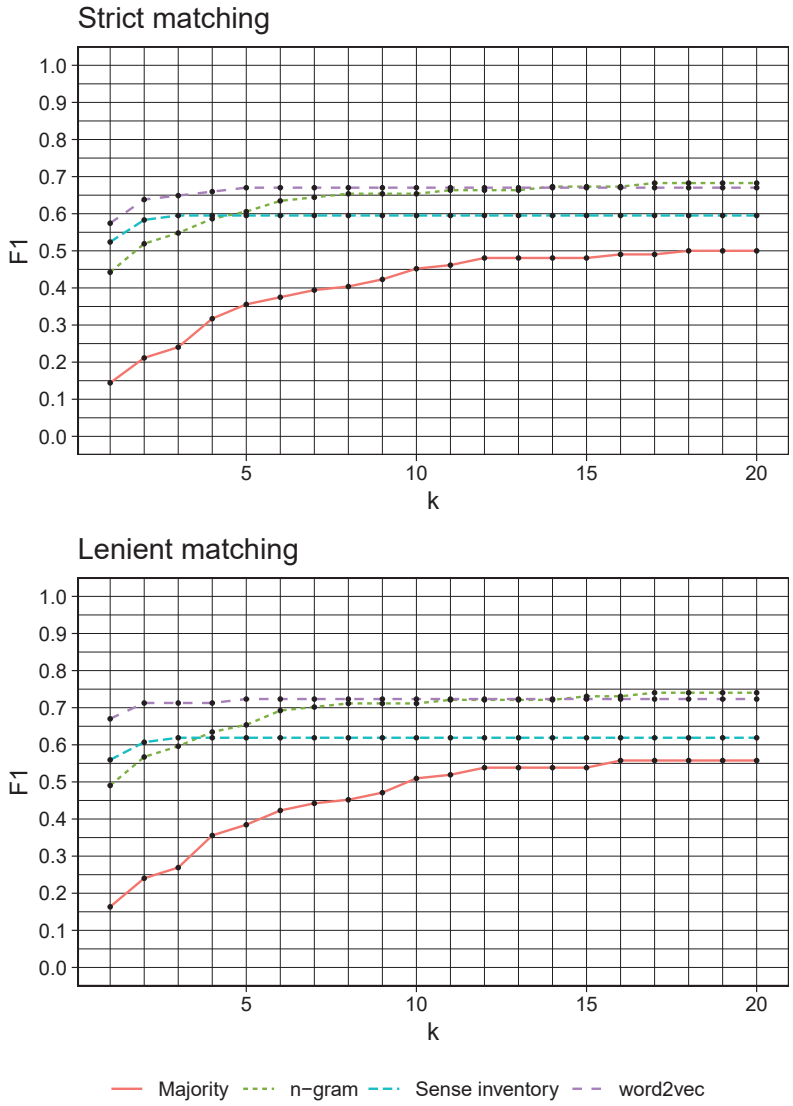

Figure 4.2: F_1-score at different ranks k for the four evaluated methods with both strict and lenient matching. Note that the expansion standard was generated at rank $k = 10$ and therefore metrics above that threshold may not be exact.

grouped several spelling variants and misspellings of "Cardiomyopathie", including the common German form starting with 'k'. Lastly, it even kept together semantic expansions such as *blood pressure* ("Blutdruck", commonly abbreviated as *RR*[14], the inventor of the cuff-based sphygmomanometer.) and *ejection fraction* ("Auswurffraktion" in German, commonly indicated by the English acronym *EF*).

QUERY	MOST SIMILAR WORDS
gutem	zufriedenstellendem; gebessertem; reduziertem; verbessertem; altersentsprechendem
Cardiomyopathie	Kardiomoypathie; Kardioymopathie; Kardiomyopatie; Kardiomypathie; Kardiomyopathie
RR	Blutdruck; Gas; Xe; Certralin; MBq Xe
EF	LVEF; Auswurffraktion; Simpson; diastolische; diast

Table 4.6: Most similar words to selected acronyms according to the WORD2VEC model, without filtering.

A deeper look at the acronyms wrongly expanded showed some common patterns, discussed below. Section 4.5.2 revisits these patterns with an overview of future work that may address several of them at once.

TRIGRAMS I generated collocations only for word pairs in the WORD2VEC approach. Experiments with collocations of three or more words (that would allow, e. g., the expansion *AIDS* → *Acquired Immune Deficiency Syndrome*) showed unexpected lower overall metrics. A closer look found no candidate expansions containing trigrams at rank $k = 10$ and thus a larger (or denser) training dataset might be needed.

LEXICALIZED ACRONYMS Some extremely common acronyms (such as *EKG* → *Elektrokardiogramm* and *LKH* → *Landeskrankenhaus*[15]) are never expanded in the collection. Although a manually curated sense inventory could easily overcome the problem, ensemble models were out of scope in the present work.

PARTIAL MATCHES Some acronyms were only partially expanded, e. g., *AINS* → *Aorteninsuff*. Note that this issue affected only strict metrics and mostly at rank $k = 1$, since lenient metrics would count such expansion as a match and a longer expansion was also commonly provided at higher ranks.

CHARACTER CHANGES Our filtering rules discarded German spelling variants, such as *CV* → *Kardioversion*. Even though dealing with character changes seems trivial at first, there is no simple normalization

14 *RR* stands for Riva-Rocci

15 LKH is the Graz University Hospital, a common acronym in the local context.

procedure that would be error-free and not require a specific evaluation[16].

SYNONYMS Our filtering rules discarded valid synonyms, such as *USKG* → *Echokardiographie* (instead of *Ultraschallkardiogramm*). Validating possible synonyms would not only require additional semantic knowledge (out of scope for the current work) but also may not be desired in an acronym expansion task where similarity and relatedness are mostly bounded by syntax.

4.5.2 *Limitations and Future Work*

Since the collection used in this study showed a low acronym ambiguity, I could not investigate methods for acronym disambiguation. A clear improvement would be to leverage the context information at runtime to better approximate the acronym vector to its correct meaning. The WORD2VEC method supports nonetheless creating sense inventories to be later used in a disambiguation task.

Given that common, lexicalized, acronyms such as *EKG* may never be expanded in a small clinical collection, the WORD2VEC and n-gram-based approaches might fail in such cases. Though a sense inventory could easily fill this gap using an ensemble approach, they may also increase ambiguity by providing expansions in different fields and thus require manual maintenance. A compromise could be reached by using specialty-specific sense inventories automatically built from, e. g., PUBMED [93].

On one hand, future work could focus on making models denser by further normalizing word variants and evaluating its impact on expansion quality. For instance, while some soundex algorithms may be able to properly normalize character changes such as k ↔ c, their output might be too degenerated to be considered a valid expansion. While normalization could be performed exclusively in the filtering step (to preserve the original spelling), this approach would possibly have a limited impact on model quality. Finally, while embedding models that exploit subword information (e. g., FASTTEXT) would in principle make models denser by mapping spelling variations nearby into the vector space, preliminary experiments showed that such misspellings were also provided as candidate expansions, an undesired side effect that would require further processing.

On the other hand, additional data could be exploited to improve model quality. If more data can be obtained, further studies could evaluate not only the impact of possible additional ambiguity but also the

16 For instance, converting all instances of "k" to "c" would lead to unacceptable forms such as "becannt" (German for "known") and "druccausgleich" (German for "pressure equalization"); conversely, converting all instances of "c" to "k" would also lead to unsatisfactory spellings such as "karotis" ("arteria carotis") and "kava" ("vena cava").

training size threshold where efficiency is satisfactory. Alternatively, one could also leverage the knowledge from larger datasets in different domains/languages with cross-language and/or cross-domain transfer learning techniques. In this scenario, e. g., the MIMIC-III dataset could be used to learn the overall structure of clinical language and have a model thereof fine-tuned in the target language.

Further work could also focus on improving, benchmarking, and automating filtering rules. Possibly the most immediate impact could be obtained by exploiting word morphemes (e. g., *EKG* → *Elektro-kardio-gramm*) — a very common linguistic phenomenon in clinical German — either at training time (by splitting long compositions) or in the filtering step (by mapping acronym characters to morpheme initials). Moreover, a more sophisticated approach would be to consider acronym expansions as a ranking problem (instead of binary filtering) so that several different signals — whose weight could be automatically learned — contribute to the candidate list ordering. In this approach, signals could include not only rational versions of the current binary rules but also the expansion tf-idf, the distance in the embedding space, as well as additional scores.

5 CLINICAL TEXT CLASSIFICATION

In this chapter I study the problem of text classification to support patient phenotyping and cohort building.

Using data from the 2018 National NLP Clinical Challenges (n2c2), I explore shallow and deep learning models associated with pre-trained word embeddings and embeddings trained on the target corpus. I show that shallow approaches like Logistic Regression (LR) and Support Vector Machine (SVM) outperform more complex approaches based on Long Short-Term Memory (LSTM) neural networks and that embeddings pre-trained on a large corpus are not better that embeddings trained on the target dataset. The rule-based approach obtained the best overall place at the n2c2 conference and our participation led to a paper published at the Journal of the American Medical Informatics Association (JAMIA).

This chapter had contributions from Amila Kugic, Markus Kreuzthaler, and Zdenko Kasáč. AK analyzed the data, implemented zoning, and presented a preliminary version of this work at the n2c2 conference. A variant of the LSTM approach was first proposed and implemented by MK. ZK analyzed false positives/negatives and gave overall clinical feedback.

5.1 INTRODUCTION

5.1.1 *Motivation*

A common secondary use of clinical data stored in computer systems is to ease cohort selection for clinical trials. While manually adding patients to a trial may induce selection bias (by preferring patients who had a more recent visit or who actively enrolled in a trial), a recruiting system can assist cohort, cross-sectional, and case-control studies [94], e. g., by drawing both case and controls subjects from large cohorts or by selecting controls for a given case sample. The usage of a non-biased population can then avoid bias in the study results [8].

However, structured information (e. g., codes from the International Classification of Diseases (ICD)) in an Electronic Health Record (EHR) system is mostly created to satisfy billing and administrative requirements [11] and thus may be biased, incomplete, or of low quality (see Section 1.1). Clinical texts typically contain a vast amount of information, but manual coding requires a long development phase and is

thus rarely used [95]. Instead, strategies for automated coding may be deployed in so-called text-based patient phenotyping [96–98].

Patient phenotyping

While automated clinical text classification (see Section 2.5) has been studied for several years (for systematic reviews, cf. Stanfill et al. [99], Shaikh et al. [100], and Christodoulou et al. [101]), novel transfer learning mechanisms might improve the efficiency of approaches based on Deep Learning (DL) (see Section 2.2), especially when dealing with small datasets like the ones typically available in the clinical domain.

5.1.2 *Related Work*

Shallow approaches

A seminal work in the field was performed by Wilcox and Hripcsak [102], who evaluated the use of different text classifiers (Naive Bayes (NB), Decision Trees, and Rule Generators) to detect six medical conditions (congestive heart failure, chronic obstructive pulmonary disease, acute bacterial pneumonia, neoplasm, pleural effusion, and pneumothorax) in 200 chest x-ray admission notes. Using the output of the MedLEE system as features, they reported that the rule generator performed best, with a sensitivity of 0.48 and a specificity of 0.99.

Jouhet et al. [103] explored both NB and SVM using three different weighting schemes — term frequency (tf), term frequency - inverse document frequency (tf-idf), and term frequency - inverse class frequency (tf-icf) — to classify 5121 pathology reports into the International Classification of Diseases for Oncology (ICD-O) and the International Agency for Research on Cancer (IARC) (International Agency for Research on Cancer) code sets. They reported best results with IARC using tf-icf: F_1-score of 0.97 for IARC categories and F_1-scores of 0.72 and 0.85 for ICD-O topography and morphology axes, respectively.

DL approaches

Exploring the superiority of DL approaches when compared to shallow methods, Lipton et al. [104] applied LSTM neural networks to classify 10 401 episodes from an intensive care unit into a terminology with 128 codes based on the ninth version of the ICD. Each episode was composed of a time series of 13 structured variables over a period ranging from 12 hours to several months. They reported that LSTM outperformed a baseline based on a multilayer perceptron with handwritten features.

Pre-trained word embeddings

Exploring the impact of pre-trained word embeddings, Yao et al. [105] investigated Convolutional Neural Network (CNN) to classify 1237 discharge summaries from the 2008 i2b2 challenge according to 15 obesity comorbidities. Their model used word embeddings (pre-trained on MIMIC) from rule-based trigger phrases and Unified Medical Language System (UMLS) concept embeddings extracted using MetaMap as features. They showed that it performed as good as simpler models using SVM and LR.

Karimi et al. [106] looked into different classification methods applied to radiology notes and corresponding ICD codes. They demonstrated that (1) a fine-tuned CNN performed similarly to shallow approaches (SVM, random forests, and LR); (2) word embeddings pre-trained on MEDLINE are better than random initialization; and (3) dynamic embeddings are better than static/frozen.

Chen et al. [45] also evaluated the impact of pre-trained embeddings from PubMed and MIMIC on clinical text classification efficiency. They showed that F_1 improved when using BioSentVec embeddings trained on both datasets, followed by PubMed and then by MIMIC alone.

Roberts [48] assessed the effect of pre-trained embeddings on LSTM and CNN models applied to two problems, namely a concept recognition and a text classification task. He verified that embeddings trained on a combination of datasets provided better results than embeddings trained on a single corpus, even if it was the target collection.

5.1.3 *2014 i2b2/UTHealth Shared Task Track 2*

Together with the University of Texas Health Science Center at Houston (UTHealth), the Informatics for Integrating Biology and the Bedside (i2b2) project organized in 2014 a shared task on clinical text classification, in which Track 2 challenged participants to identify eight risk factors for heart disease (diabetes mellitus, cardiovascular disease, hypertension, hyperlipidemia, obesity, smoking, family history, and medication).

Stubbs et al. [107] reported that teams explored several different strategies, but since most of them were hybrid, there was no consensus on what exactly performed better. They also discussed that those groups obtained the lowest metrics on cardiovascular disease and on risk factors that depended on data encoded in pseudo-tables (i. e., tables manually created in the text using, e. g., tabulation marks).

Kotfila and Uzuner [108] presented a systematic analysis of the impact of weighting schemes, feature spaces, kernels, and training size on the efficiency of an SVM classifier. They could not find statistically significant differences between (1) lower-cased alphabetic tokens and semantically enriched tokens using MetaMap; (2) tf-idf and tf; (3) linear and radial kernels; and (4) small and larger training datasets.

Roberts et al. [109] re-annotated two-thirds of the training data at a finer level guided by a hand-crafted lexicon and then trained SVM classifiers using the newly labeled data. They pre-processed documents using ConText to identify negation and modality markers, as well as section headers. With the proposed approach, they reported precision of 0.90, recall of 0.96, and F_1 of 0.93, the best metrics in the shared task.

In 2018, the shared task was renamed to n2c2 and started to be organized by the Department of Medical Informatics of the Harvard Medical School.

5.1.4 *Research Problem*

Given recent works showing contradicting results on shallow and deep learning approaches for text classification (see also Joulin et al. [63] discussed in Section 2.5) and that pre-trained embeddings with subword information were just recently released for the clinical domain [45], I decided to use data from the 2018 n2c2 to revisit the topic and fill the experimentation gap with additional data.

I thus aim to assess the impact of pre-trained embeddings for text classification using a small clinical collection, both when using shallow and deep learning methods. I hypothesize that shallow methods obtain better results than deep learning approaches and that pre-trained embeddings improve results in a small clinical dataset.

5.2 MATERIALS

5.2.1 *Dataset*

The 2018 n2c2[1] Shared Task Track 1 reused the corpus from the 2014 i2b2/UTHealth Shared Task Track 2. Participants were asked to develop a text classifier that determines whether a patient meets or does not meet a set of selection criteria for a clinical trial. We received in advance a training dataset with 887 clinical narratives from 202 patients and later a test dataset with 377 narratives from 86 extra patients (see Table 5.1).

	TRAINING	TEST
Ratio	~70 %	~30 %
Patients	202	86
Narratives	887	377
Words	547 625	230 221
Annotations	2626	1118

Table 5.1: Overview of the dataset splits provided in the 2018 n2c2 Shared Task Track 2.

Figure 5.1 depicts an excerpt of a sample document and the annotations provided. Each file represented a single patient with up to five narratives corresponding to time-annotated visits.

Out of the thirteen selection criteria (see Table 5.2), six were balanced in the dataset (one class with at least one-third of patients),

1 https://n2c2.dbmi.hms.harvard.edu

```
<?xml version="1.0" encoding="UTF-8" ?>
<PatientMatching>
<TEXT><![CDATA[
****************************************************************
Record date: 2069-11-02

(...)

PAST MEDICAL HISTORY:  insulin-dependent diabetes since 25yo, retinal
    neuropathy, asthma. Obesity.

MI treated here previously.  Ischemia.

SOCIAL/FAMILY HISTORY: No alcohol.  Smoked in the past, quit 10 years ago
    . Family history

of ischemia and CAD

MEDICATIONS:
1. Provigil
2. Atenolol
3. Ativan.
4. Glucophage 850 mg t.i.d.
5. Humulin 15 units at night.
6. Folate.
7. Metoprolol.
8. Cardia.
9. Vitamin E.
10. Coated aspirin.

Recommended full cardiac evalution (sic); possible need for stent.
    Patient opted to return following day.

****************************************************************
]]></TEXT>
<TAGS>
<ABDOMINAL met="not met" />
<ADVANCED-CAD met="met" />
<ALCOHOL-ABUSE met="not met" />
<ASP-FOR-MI met="met" />
<CREATININE met="not met" />
<DIETSUPP-2MOS met="met" />
<DRUG-ABUSE met="not met" />
<ENGLISH met="met" />
<HBA1C met="met" />
<KETO-1YR met="not met" />
<MAJOR-DIABETES met="met" />
<MAKES-DECISIONS met="met" />
<MI-6MOS met="met" />
</TAGS>
</PatientMatching>
```

Figure 5.1: A sample document in the 2018 n2c2 Shared Task Track 2 (edited for brevity).

six imbalanced (one class with less than 10% of the patients), and one semi-balanced (one class with between 10% and one-third of patients). Moreover, two criteria (HBA1C and CREATININE) were value-dependent, i. e., they were characterized by a range of values.

CRITERION	BALANCE	DESCRIPTION
ABDOMINAL	Balanced	History of intra-abdominal surgery.
ADVANCED-CAD	Balanced	Presence of advanced cardiovascular disease.
ALCOHOL-ABUSE	Imbalanced	Current alcohol use over recommended limits.
ASP-FOR-MI	Semi-balanced	Use of aspirin to prevent myocardial infarction.
CREATININE	Balanced	Serum creatinine above the normal limit.
DIETSUPP-2MOS	Balanced	Use of dietary supplements (last two months).
DRUG-ABUSE	Imbalanced	Drug abuse.
ENGLISH	Imbalanced	The patient can speak English.
HBA1C	Balanced	Glycated hemoglobin between 6.5% and 9.5%.
KETO-1YR	Imbalanced	Ketoacidosis in the last year.
MAJOR-DIABETES	Balanced	Major complication due to diabetes.
MAKES-DECISIONS	Imbalanced	The patient can make decisions by himself.
MI-6MOS	Imbalanced	Myocardial infarction in the last six months.

Table 5.2: Overview of the thirteen classification criteria. Reproduced from Oleynik et al. [1] with permission of the Oxford University Press (original work distributed under the terms of the CC-BY-NC license).

5.2.2 *Evaluation Metrics*

The official evaluation metrics[2] were precision, recall, and F_1-score per class ("met" and "not met") and selection criteria on the test set (see Section 3.2). Overall F_1-score for a given criterion was the simple mean of F_1-scores for the two classes. Additionally, I defined "met" and "not met" as positive and negative outcomes, respectively, in order to calculate accuracy per criterion. Finally, macro and micro-averaged metrics were calculated over the thirteen criteria. The official metric used to rank participating teams was overall micro F_1-score.

5.2.3 *Pre-trained Word Embeddings*

I experimented with both BIOWORDVEC embeddings pre-trained in PUBMED and MIMIC-III [45] and embeddings trained in the target dataset (n2c2). For the latter, I used FASTTEXT with the same hyperparameters: a window size of 20, a maximum n-gram length of 6, a learning rate of 0.05, and a negative sample size of 10.

2 Metrics were calculated by the official evaluation script available at https://github.com/filannim/2018_n2c2_evaluation_scripts.

5.3 METHODS

Table 5.3 shows an overview of the proposed methods. We implemented five orthogonal (non hybrid) methods, namely a baseline set to a majority classifier, a rule-based classifier, two shallow classifiers (SVM and LR), and a deep LSTM neural network. For the methods using input pre-training (LR and LSTM), I explored both self-trained (SELF) and pre-trained word embeddings (PRE). I also performed McNemar's tests (see Section 3.3.2) comparing each method to the baseline to check whether any differences were statistically significant, with $\alpha = 0.05$. Our source code is available at GitHub[3] under the Apache License V2.

ACRONYM	CLASSIFICATION METHOD	WORD EMBEDDINGS
Baseline	Majority	N/A
RBC	Rule-based	N/A
SVM	Support Vector Machine	N/A
SELF-LR	Logistic Regression	Self-trained
PRE-LR		Pre-trained
SELF-LSTM	Long Short-Term Memory	Self-trained
PRE-LSTM		Pre-trained

Table 5.3: Overview of the proposed methods. Reproduced from Oleynik et al. [1] with permission of the Oxford University Press (original work distributed under the terms of the CC-BY-NC license).

5.3.1 *Preprocessing*

We pre-processed the documents using the techniques described in Section 2.3.1, viz. general cleaning, sentence detection, tokenization, lowercasing, stopword, and punctuation removal.

Given that DIETSUPP-2MOS, KETO-1YR, and MI-6MOS were time-dependent criteria, we split each patient text into visits (text zoning), clearly separated by asterisks in the raw data. Moreover, we parsed the record date on the first line, so that we could properly restrict the input data for the aforementioned criteria. This time restriction improved training results for both SVM and the Rule-Based Classifier (RBC).

5.3.2 *Rule-based Classifier*

The rule-based approach was based on 111 positive/negative markers (such as "elevated creatinine") and two regular expressions for value extraction. Values extracted with regular expressions for the criteria

3 https://github.com/bst-mug/n2c2

HBA1C and CREATININE were compared to thresholds obtained from the training set, namely $6.5 \leqslant \text{hba1c} \leqslant 9.5$ and $1.4 < \text{creatinine} < 10$.

Except for the criteria ENGLISH and MAKES-DECISIONS (the most imbalanced criteria towards "met"), all markers were positive, i. e., they triggered the class "met" when found. I opted to fall back to a majority ("not met") approach for the KETO-1YR criterion due to the lack of positive examples. On the other hand, ADVANCED-CAD was the most complex criterion, with 28 textual markers to classify the four "advanced" conditions specified in the annotation guidelines, on top of which counting rules would make the final classification.

After successive rounds of false positive/negative analysis, I found that, in a few cases, important markers would (a) occur in a negated context (such as "denies ischemia" and "no retinopathy"); (b) refer to allergies ("allergy to aspirin" and "aspirin: avoid"); (c) refer to past history ("STEMI in 2008"); or (d) refer to family history ("FH with NSTEMI"). I then enriched such text markers with negative "lookaround" regular expressions.

5.3.3 *Support Vector Machine*

Support Vector Machine is a discriminative linear classifier that learns a decision hyperplane that maximizes the decision margin between the annotated samples. Given the decision hyperplane with normal $\vec{w}$ and the intercept term b learned during training, a SVM can classify a new input vector $\vec{x}$ following Equation 5.1.

$$f(\vec{x}) = \text{sign}(\vec{w}^T \vec{x} + b) \tag{5.1}$$

For text classification, SVM is commonly associated with a bag-of-words representation of the input documents. Text is then either represented with a tf-idf weighting scheme or word embeddings, as described in Section 2.3.2. Even though SVM can be expanded with the so-called "kernel trick" to change the decision surface to, e. g., a polynomial or Gaussian function, a linear kernel associated with a bag-of-words representation using tf-idf is known to be a strong baseline [61].

I used Weka[4] 3.8.2 to train the model with a Java wrapper for the LIBSVM[5] 2.82 library.

5.3.4 *Logistic Regression*

I also experimented with a multinomial LR[6] strategy, which can be implemented as a single-layer feed-forward neural network (*perceptron*

4 https://www.cs.waikato.ac.nz/ml/weka/
5 https://www.csie.ntu.edu.tw/~cjlin/libsvm/
6 Also known as Maximum Entropy (MaxEnt).

network) with logistic activation functions. The logistic function is defined as in Equation 5.2, where $\vec{w}$ is the vector $[w_0, w_1]$ of weights learned during the training process. *Perceptron network*

$$\text{Logistic}(\vec{x}) = \frac{1}{1 + e^{-\vec{w}\vec{x}}} \tag{5.2}$$

I applied multinomial LR using the default setting of FASTTEXT[7] 0.2.0, which represents documents as the average of word vectors contained therein. I initialized these vectors with either pre-trained clinical embeddings from BIOWORDVEC (PRE-LR) or clinical embeddings trained in the target collection (SELF-LR). I trained FASTTEXT during 100 epochs with a learning rate of 0.5, a window size of 5 words and the Cross Entropy (XENT) loss function, which were found to be optimal in the training set.

5.3.5 *Long Short-Term Memory*

First proposed by Hochreiter and Schmidhuber [110], LSTM networks are a special kind of Recurrent Neural Network (RNN) containing a self-loop (apart from the outer loop present in every RNN) that can preserve state. Additionally, the weight of this gated unit is not fixed but controlled by another hidden unit, thus allowing the network to dynamically learn sub-sequences [25]. LSTM is commonly used to model time-series events [104], but has been gaining popularity in the natural language area [111, 112].

We represented text either as a sequence of BIOWORDVEC word embeddings (PRE-LSTM) or embeddings trained in the target collection (SELF-LSTM). We fed the frozen embeddings through a Neural Network (NN) containing a single layer of LSTM cells — each with a hidden state vector of size 64 — and a fully-connected output layer with the sigmoid activation function and Cross Entropy (XENT) loss. We modeled our network using the DEEPLEARNING4J[8] 0.9.1 framework and trained it using the gradient-based Adam [113] function during 25 epochs with a learning rate of 0.02 and dropout of 0.5 (hyperparameters chosen empirically).

5.4 RESULTS

Figure 5.2 and Figure 5.3, respectively, present F_1 and accuracy of the proposed methods and the baseline. Table 5.4 summarizes the findings by providing overall metrics of the different methods when compared to the baseline. Detailed results per target class are shown in Section A.1. A detailed overview of the results per method is presented next.

7 https://fasttext.cc/
8 https://deeplearning4j.org

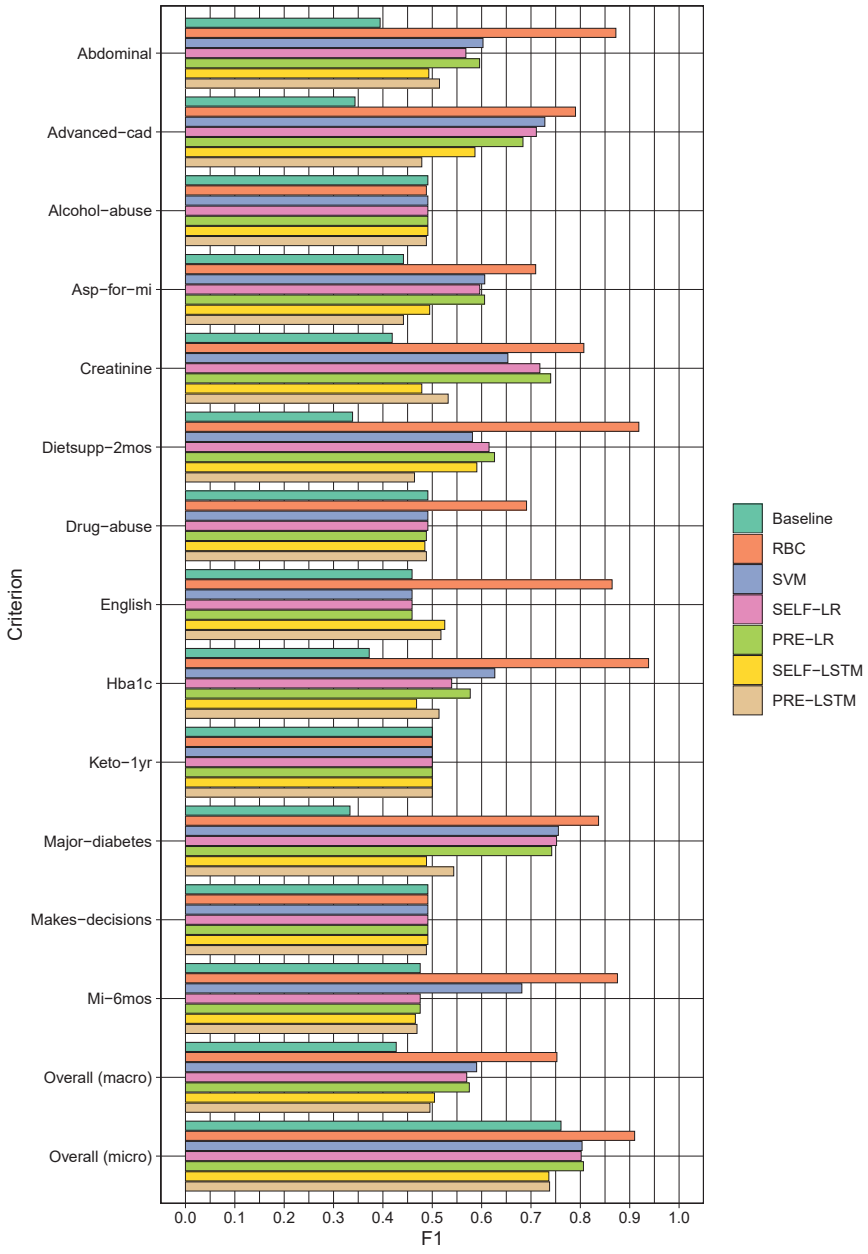

Figure 5.2: Overall F_1 score per criterion on the test set.

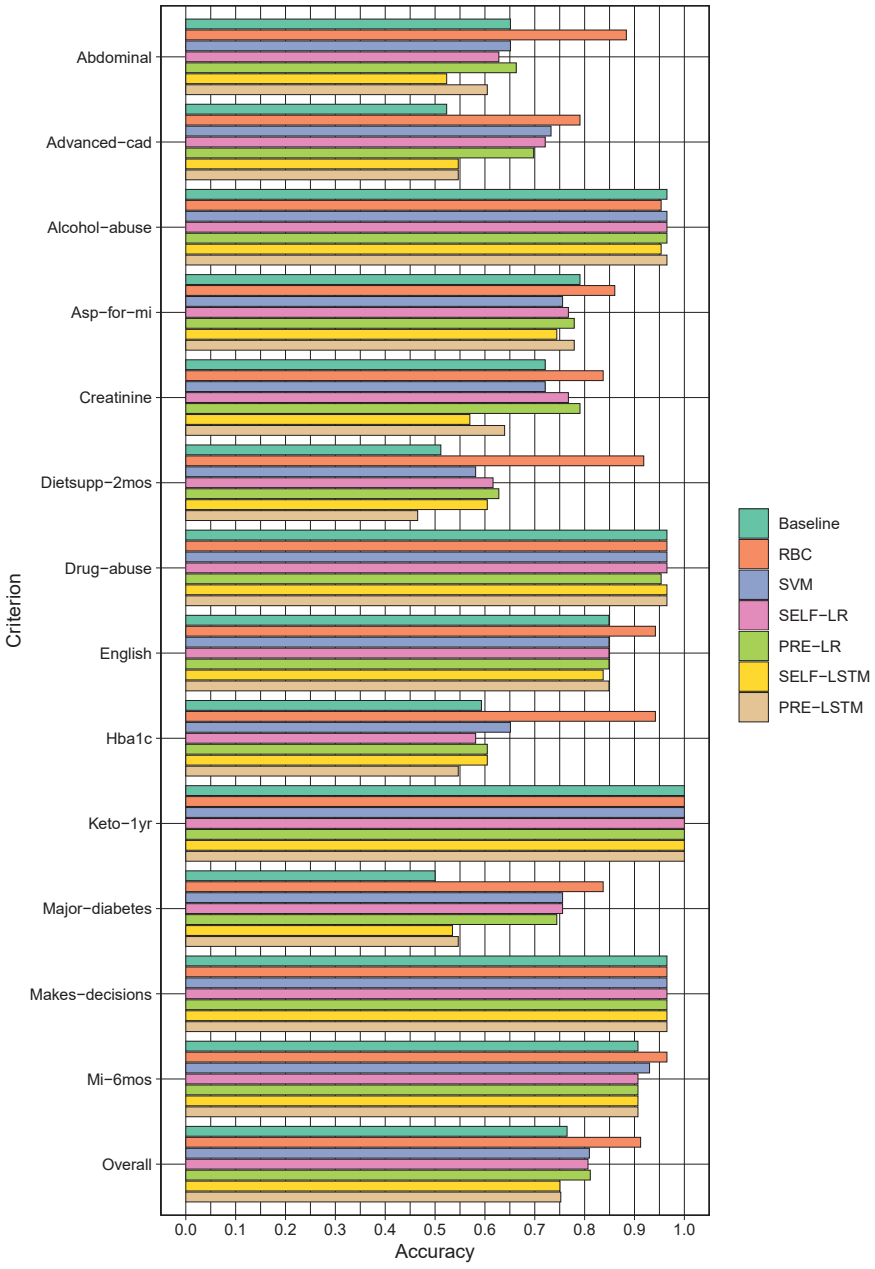

Figure 5.3: Overall accuracy per criterion on the test set.

METHOD	TP	FP	TN	FN	C	F_1	A
Baseline	356	160	499	103	855	0.7608	0.7648
RBC	419	58	601	40	1020	0.9100	0.9123
SVM	355	109	550	104	905	0.8035	0.8095
SELF-LR	361	118	541	98	902	0.8017	0.8068
PRE-LR	364	116	543	95	907	0.8063	0.8113
SELF-LSTM	347	167	492	112	839	0.7362	0.7504
PRE-LSTM	338	156	503	121	841	0.7377	0.7522

Table 5.4: Overall results on the test set. TP = true positives, FP = false positives, TN = true negatives, FN = false negatives, C = correct.

BASELINE Individual F_1-scores did not exceed 0.5000 as expected, but accuracy values reached high values in some criteria due to class imbalance. This also reflected on a high overall micro F_1-score of 0.7608 and an accuracy of 0.7648.

RULE-BASED CLASSIFIER Compared to the baseline, RBC showed higher F_1-score and accuracy on every criterion except ALCOHOL-ABUSE due to a single false positive caused by a missed negation. RBC improved the absolute metrics the most on the DIETSUPP-2MOS criterion ($\Delta F_1 = +0.5800$ and $\Delta A = +0.4070$). The lowest F_1-scores were obtained on imbalanced criteria such as ALCOHOL-ABUSE, but they did not impact the overall micro F_1-score due to micro-averaging. Conversely, the criteria ADVANCED-CAD, CREATININE, and MAJOR-DIABETES showed the lowest values of accuracy.

SUPPORT VECTOR MACHINE Compared to the baseline, SVM showed equal or better values of F_1-score and accuracy for all criteria except ASP-FOR-MI, a criterion that showed an increase on false negatives (caused by aspirin being a common drug and thus having a lower idf) and therefore lower accuracy. We obtained the lowest values of F_1-score on the imbalanced criteria ENGLISH, ALCOHOL-ABUSE, and MAKES-DECISIONS. Conversely, the criteria DIETSUPP-2MOS and ABDOMINAL presented the lowest values of accuracy (due to the higher intrinsic variance compared to the amount of available training data), followed by the criteria HBA1C and CREATININE, selection criteria whose text indicators are number-based (e. g., "HBA1C 6.5") and thus cannot be captured by a bag-of-words approach.

LOGISTIC REGRESSION Considering individual F_1-scores, both SELF-LR and PRE-LR had equal or better results than the baseline for all criteria except DRUG-ABUSE for PRE-LR. With respect to accuracy, SELF-LR (PRE-LR) had equal or better results than the baseline for all criteria except ABDOMINAL, ASP-FOR-MI, and HBA1C (ASP-FOR-MI and DRUG-ABUSE). Both SELF-LR and PRE-LR obtained the

lowest F_1-score on the imbalanced criterion ENGLISH; conversely, SELF-LR (PRE-LR) had the lowest accuracy on the criterion ABDOMINAL (DIETSUPP-2MOS).

LONG SHORT-TERM MEMORY The SELF-LSTM (PRE-LSTM) classifier had individual F_1-scores equal or better than the baseline for every criterion except DRUG-ABUSE and MI-6MOS (ALCOHOL-ABUSE, DRUG-ABUSE, MAKES-DECISIONS, and MI-6MOS). However, SELF-LSTM (PRE-LSTM) showed lower or equal values of accuracy for most criteria and improved over the baseline only for the criteria ADVANCED-CAD, DIETSUPP-2MOS, HBA1C, and MAJOR-DIABETES (ADVANCED-CAD and MAJOR-DIABETES). The criteria MI-6MOS and ASP-FOR-MI presented the lowest F_1-score for SELF-LSTM and PRE-LSTM, respectively. Conversely, the criteria ABDOMINAL and DIETSUPP-2MOS had the lowest accuracy for SELF-LSTM and PRE-LSTM, respectively.

5.5 DISCUSSION

The results presented in the previous section showed that the rule-based classifier obtained the highest results and was significantly better than the baseline set to a majority approach (RBC: $p < 0.001$); shallow methods presented the second highest metrics and were still significantly better than the baseline ($p < 0.001$); and deep learning results were not different than the baseline (SELF-LSTM: $p > 0.05$, PRE-LSTM: $p > 0.05$). Moreover, I could not show a difference in classification efficiency between the usage of pre-trained embeddings and embeddings trained on the target data neither for shallow ($p > 0.05$) nor for deep learning methods ($p > 0.05$).

5.5.1 *Error Analysis*

Apart from the higher efficiency, the rule-based approach also made the interpretation of the results easier. It is thus useful to show some of the issues in properly classifying clinical text, which may also affect machine learning strategies; therefore, I provide below a breakdown of issues per criterion.

ABDOMINAL Even though "renal transplant" is a strong and common marker for intra-abdominal surgeries (37 occurrences in the training set), I found that its meaning was sometimes inverted by nearby words. For instance, "renal transplant evaluation", "waiting for donor for a renal transplant", and "postponing her renal transplant" do not denote a real case of transplant.

ADVANCED-CAD Even though "chest pain" is a common synonym for angina, it is commonly used in negated sentences such as "de-

nies ... chest pain", "negative for ... chest pain", and "chest pain free". Using it as a marker generates many false positives, whereas the term "angina" is mostly used with a positive polarity and preferred when describing this condition.

ALCOHOL-ABUSE I missed the marker "vodka" not seen during training; however, adding all terms from a terminology could lead to false positives due to negations. I could also have parsed the number of beers/drink per day, a non-trivial task though.

ASP-FOR-MI I found several false positives with patients who took aspirin for other reasons than for myocardial infarction prevention. Detecting such cases is a challenging task because the indication is seldom mentioned in medication lists from clinical documents.

CREATININE I saw in the training set that 1.4 mg/dL in the serum seems to be a common threshold of normality for this laboratory test, but I found conflicting examples for values closer to that. Also, results obtained via emergency (STAT) laboratories may have a higher cut-off to avoid false positives.

DIETSUPP-2MOS Both "calcium" and "magnesium" are minerals that occur in blood sample measurements, but they are also prescribed as a dietary supplement. Additionally, both are followed by digits, which makes them hard to disambiguate.

DRUG-ABUSE The choice for explicit markers such as "cocaine" instead of the more general "drug abuse" seemed to be a good compromise for an unbalanced class, as the former is seldom negated, while the latter is more commonly used in negation lists.

ENGLISH The use of more general markers such as "interpreter" (instead of "with interpreter") and "translat*" (instead of "translated") could have prevented me from missing cases such as "acted as her interpreter" and "translating from Columbian (sic)".

HBA1C Some documents were misclassified due to: (1) capturing the value only when a decimal place was present; (2) matching only "A1C" and not, e. g., "hemoglobin of 6.5"; (3) misspellings with a lowercase L instead of the digit 1[9], e. g., "HB Alc" and "HgAlC".

KETO-1YR There was only one positive example of ketoacidosis in the training set and therefore I obtained perfect results in the test set.

9 In old typewriters, there was not a key for the digit 1 and users would type using a lowercase L instead.

MAJOR-DIABETES I did not use the marker "ESRD" (end-stage renal disease) because it was secondary to hypertension (and not hepatogenous diabetes) in some cases. Detecting such differences, although hard, would have helped to improve results.

MAKES-DECISIONS I opted to match only documents mentioning "severe dementia", as "dementia" is used in the family history (e. g., "Father had dementia"); however, I missed some cases due to this decision.

MI-6MOS As I preferred to use the very precise "STEMI" marker during training, I missed at least one case where "NON ST elev MI" is mentioned instead. However, matching "MI" only would have led to an increase in false positives, due to it also being used to describe a non-specific myocardial infarction in the family history, as well as being the acronym for the U.S. state of Michigan.

Overall, most false negatives could be explained by expressions used in the test set that never occurred in the training set, an issue that could be addressed by additional data. Conversely, false positives were mostly caused by the marker being changed due to negation, polarity, or presence in the family history.

5.5.2 *Shared Task Results*

When compared to all 46 participating teams at the 2018 n2c2 Shared Task Track 1, the rule-based classifier obtained the best overall metrics (see Table 5.5) [114]. Except for the deep learning approach, our proposed methods sit above the average (0.7989), but below the median (0.8227). The winning rule-based approach is above the median by 0.75 times the standard deviation (0.1161).

It is interesting to observe that four out of the ten best teams used rule-based approaches, while the other six explored hybrid approaches that incorporated rules somehow. For instance, Vydiswaran et al. [115] explored a similar, but more complex, approach based on (1) vocabulary lists extracted from domain websites; (2) regular expressions for value-based criteria; and (3) hand-crafted rules applied to sections and concepts identified by off-the-shelf Natural Language Processing (NLP) tools. Tannier et al. [116] employed an almost identical criterion-dependent approach based on (1) and (2), but also augmented the training set with notes from MIMIC using hand-crafted rules observed from the n2c2 training dataset and then trained LR classifiers. Likewise, Chen et al. [117] exploited lexical cues based on term lists and syntactic rules leveraging negation, certainty, and past/family histories.

TEAM	METHOD	DESCRIPTION	F_1
Medical University of Graz [1]	**Rule-based**	**Regular expressions**	**0.9100**
University of Michigan [115]	Hybrid	Rules + external resources	0.9075
Sorbonne Université [116]	Hybrid	Rules + external resources + ML	0.9069
Med Data Quest [117]	Rule-based	"Bottom-up", modular	0.9028
Cincinnati Children's Hospital Medical Center [118]	Hybrid	Rules + external resources + ML	0.9026
Arizona State University [119]	Hybrid	CLAMP + rules + SVM	0.9003
University of New South Wales, National Cancer Institute [120]	Rule-based	Physician assistance	0.8913
Harbin Institute of Technology [121]	Hybrid	Rules + CNN + LSTM	0.8855
University of Utah [122]	Rule-based	Trie-based hash rule	0.8837
National Taitung, Taipei Medical, University of New South Wales	Hybrid	SVM with polynomial kernel	0.8765

Table 5.5: Top overall systems in the 2018 n2c2 Track 1.

5.5.3 *Limitations and Future Work*

Although negation and family history contexts are common phenomena in medical reports, I addressed them only partially in the rule-based approach based on examples seen in the training data. A more comprehensive approach would have been to exploit available tools such as NegEx and ConText and to apply them to the three presented methods.

Even though I tried to keep the rule-based approach reusable and maintainable by creating a minimal set of markers, further work is required to evaluate its application in other institutions. On the other hand, the SVM and LR classifiers are domain and language independent, which makes them easier to be reused in other institutions.

The shallow approaches based on SVM and LR are limited in other ways though, including: (1) they assume word independence and thus cannot distinguish, e. g., "vitamin A" from "A vitamin"; (2) they do not capture value-based data needed for the HBA1C and CREATININE criteria; Even though I could have mitigated (2) with different approaches, using a single bag-of-bigrams could have helped (1); however, training this model would need more data due to the larger dimensionality.

Moreover, the SVM classifier was restricted to the top 1000 most common words as features. In the future, I would like to explore information gain or the chi-square statistical test to better prune the feature space without a negative impact on efficiency. This is especially important when processing a small dataset like n2c2's, in which a larger number of features might lead to overfitting. It is interesting to observe, however, that the LR approach using word embeddings might have

automatically benefited from the lower dimensionality without any further processing and a smaller loss of information.

Even though we applied basic text normalization steps, I did not explore the impact of short form resolution (see Chapter 4) or spell checking in our models because I did not identify such phenomena in the training collection. Nonetheless, the models based on word embeddings with subword information may have been able to capture such variants by generating word vectors clustered together in the vector space.

Some data points such as laboratory values were also hidden in pseudo-tables encoded as text. Since such tables do not always follow the same spacing or structure, I did not attempt to retrieve such information. I noticed, however, that physicians would normally discuss abnormal values in the surrounding text and thus provide a useful summary comprehensible not only by humans but also by machines. Structured data may also be available from other data sources in a real clinical setting.

Future work might also employ sentence parsing, thus allowing the understanding of sentences such as "BUN and creatinine were 32 and 1.2" and "creatinine fell from 2.2 to 1.6". Nevertheless, models for advanced clinical natural language processing still lack efficiency when compared to models trained in the general domain.

As stated before, we explored only orthogonal, non-hybrid, strategies to increase comparability among methods and thus improve consensus on optimal solutions for this class of problems [107]. However, hybrid solutions may be recommended for a real scenario, either via (a) stacking methods, e. g. SVM trained on top of features extracted with rules [116, 121]; (b) an ensemble of methods linearly combining predictions provided by each algorithm; or (c) a mixed approach with a specific method for each criterion, e. g., machine learning for balanced and complex classes and rules for imbalanced or value-based classes [115, 116].

6 BIOMEDICAL INFORMATION RETRIEVAL

In this chapter I explore word embeddings for query expansion of documents from biomedical collections and compare them to traditional approaches based on rules and terminologies.

Using data from the first two editions of the Text REtrieval Conference - Precision Medicine (TREC-PM), we proposed a method that increases recall without impacting precision. Using this method, I show that word embeddings can be used to increase recall when terminologies are not available and that the benefit of query expansion is stronger in a small dataset. A system containing some of this thesis' contributions led to the best overall place in the 2019 edition of TREC-PM.

This chapter had contributions from Ariane Morassi Sasso, Erik Faessler, Jan Philipp Sachs, Pablo López-García, Stefan Schulz, and Zdenko Kasáč. The experimental framework received contributions from AMS, EF, and PLG. The Elasticsearch (ES) index was initially created by PLG and maintained by EF. The idea of using `dis_max` queries was first discussed with AMS and EF in July 2018. Internal reference standards used to derive rules and keyword boosters were partially annotated by AMS, JPS, SS, and ZK. The keyword boosters were independently evaluated by AMS. A preliminary version of query expansion using terminologies was first proposed by PLG. Different versions of this work were presented at the TREC-PM conference by PLG (2017), AMS (2018), and EF (2019).

6.1 INTRODUCTION

6.1.1 *Motivation*

The biomedical domain is especially prone to linguistic phenomena such as synonymy, hypernymy, and meronymy. For instance, (1) "cholangiocarcinoma" is a kind of carcinoma (hypernymy), a cancer of the biliary duct (synonymy), the latter a body structure that carries bile from the liver and gallbladder to the duodenum, by itself the first section of the small intestine (meronymy); (2) colon, rectum, and anal cancers are normally referred together as "colorectal cancers" (hypernymy) or "bowel cancer" (synonymy) for being the final parts of the intestine (meronymy); (3) both "ERBB2" and "HER2" refer to the same gene (synonymy), whose acronyms (a particular case of synonymy, see Chapter 4) refer to "erythroblastic oncogene B 2" and "human epidermal growth factor receptor 2", respectively.

Reference terminology

Interface terminology

Some of these mappings among concepts have traditionally been manually maintained in reference terminologies like SNOMED CT. However, terminological descriptions may not reflect real-life term usage and therefore interface terminologies have been proposed to provide additional familiar terms to the users [123, 124]. For instance, an interface terminology may contain the mapping between the terms "heart attack" and "myocardial infarction" to "ischemic injury and necrosis of heart muscle cells resulting from absent or diminished blood flow in a coronary artery" [123].

While terminologies need to be constantly curated, approaches leveraging word co-occurrence statistics such as Latent Semantic Analysis (LSA) and Word Embeddings (WE) may provide automatic updates only based on newer corpus snapshots. In the biomedical domain, WE pre-trained on the large PubMed collection (see Section 3.1.2) may be a rich source for semantically similar terms and therefore proxy the requirements of a manually curated interface terminology.

In an Information Retrieval (IR) scenario (see Section 2.6), linguistic variations play an important role because a typically short user query must be expanded with the appropriate synonyms and, depending on the query, also hypernyms and meronyms. Query expansion improves recall but may however lead to a drop in precision (owing to the precision-recall tradeoff, see Section 2.6) if an unbalanced or wrongly expanded query leads to the retrieval of irrelevant documents. The main focus of this work is, therefore, to use novel strategies to improve recall in the biomedical domain via query expansion without a corresponding drop in precision.

6.1.2 *Related Work*

TREC Ad Hoc

Since 1992, the National Institute of Standards and Technology (NIST) has organized the Text REtrieval Conference (TREC) to improve the field of IR. Even though several tracks have been organized each year, the TREC Ad Hoc track held until 1999 provided the most known test set, a collection of 1.89 million newswire articles [59, Section 8.2].

Roy et al. [125] explored word embeddings for query expansion using TREC Ad Hoc data. They showed that word embeddings improved over a non-expanded baseline, but performed worse than a co-occurrence based method. These results were later corroborated by Kuzi et al. [126] in a similar dataset.

TRECMed

More recently, tracks in the biomedical domain have been proposed. From 2011 to 2012, it offered the TRECMed track, for which participants were asked to rank 95 702 medical records provided by the University of Pittsburgh's BLULab NLP Repository given a textual topic describing a cohort [127–129]. Unfortunately, the data is only available to previous participants and no further experiments could be performed.

From 2014 to 2016, NIST offered the Text REtrieval Conference - Clinical Decision Support (TREC-CDS) track with the task of ranking documents from PubMed Central (PMC) that provide support for the diagnosis, test, or treatment for patients represented by short excerpts of their de-identified clinical data [130–132]. *TREC-CDS*

Goodwin and Harabagiu [133] proposed using word embeddings in the context of query expansion for the 2014 edition of the TREC-CDS. They used the top-20 expansions provided by WORD2VEC trained on the target dataset and compared them to expansions based on classical Latent Dirichlet Allocation (LDA) and terminologies. They reported that systems based on LDA and WORD2VEC performed the weakest.

Nguyen et al. [134] reviewed the impact of query expansion in the 2016 edition of the TREC-CDS challenge. They showed that three participating teams used word embeddings trained either on WIKIPEDIA or PMC and eight teams expanded words using the Unified Medical Language System (UMLS) or the Medical Subject Headings (MeSH). Since the results of most teams were inconclusive (because the submitted runs were a combination of several features, with often unreported parameters), they attempted to reproduce the experiments by training word embeddings on a combination of WIKIPEDIA and MEDLINE, as well as comparing them to UMLS-based expansion. Compared to a baseline run, they obtained better results in only two of four metrics when using either word embeddings or a thesaurus. The two mentioned methods had similar results.

Starting in 2017, NIST has organized the TREC-PM track and asked participants to retrieve, for a given set of input topics representing patients, relevant documents from two given collections, namely Biomedical Abstracts (BAs) from PUBMED and Clinical Trials (CTs) from CLINICALTRIALS.GOV [135, 136]. TREC-PM differs from TREC-CDS by the presence of structured topics, which may ease query expansion with a reduced noise. *TREC-PM*

During the 2017 edition, two participating teams explored the usage of word embeddings for query expansion. The team from Nguyen et al. [137] reused the embeddings trained for TREC-CDS (see above) and expanded terms from the topic with the three most similar embeddings; however, they obtained inconclusive results due to mixed experiments. Eghlidi et al. [138] trained embeddings on WIKIPEDIA articles mentioning a curated list of genes from HGNC and similarly expanded topics with the three most similar words; results showed little improvement due to judgments being based on exact matches.

Two different participating teams explored this approach in 2018. Nishani et al. [139] expanded topics with the five most similar word vectors obtained from embeddings trained on a union set of WIKIPEDIA, PMC, and WIKIPEDIA; results were again not promising. Likewise, Baruah et al. [140] used the top-most similar word provided by embeddings trained on PUBMED and compared it to a thesaurus-based

expansion method; embeddings showed only a mild effect over a baseline, while the results with thesauri were inconclusive due to mixed experiments.

Nevertheless, there is some evidence that more specific embeddings might have helped the participating teams. Diaz et al. [141] explored the impact of training word embeddings in a topic-based context (local) as opposed to training in the full dataset (global). They showed that locally-trained embeddings outperform global embeddings even if the training dataset is not the same as the retrieval dataset. Conversely, when training embeddings on the target retrieval dataset, topic-based embeddings seemed to outperform global embeddings, but only by a small margin. In a similar study, Rattinger et al. [142] confirmed the previous findings using collections from the Conference and Labs of the Evaluation Forum (CLEF) and Association for Computational Linguistics (ACL).

6.1.3 *Research Problem*

On the one hand, the biomedical domain is prone to synonymy, hyponymy, and hypernymy, with terminology interactions being constantly updated owing to scientific progress. On the other hand, there is an increasing demand to provide precise and updated results tailored to a specific patient in the new era of precision medicine. Given inconclusive results on the usage of word embeddings for global query expansion on IR, I propose to use data from the 2017 and 2018 TREC-PM competitions to cast additional light on the topic. I hypothesize that (a) queries can be expanded without a drop in precision, and (b) word embeddings can be employed to substitute rules and terminologies for query expansion.

6.2 MATERIALS

6.2.1 *Dataset*

We participated in the 2017 to 2019 editions of the Precision Medicine track[1] of the TREC challenge organized by NIST and used the data from the first two years in this study[2]. Figure 6.1 provides an overview of the task. The challenge was to retrieve, for a given set of input topics representing patients, relevant documents from two given collections, namely biomedical abstracts from PubMed and clinical trials from ClinicalTrials.gov.

1 http://www.trec-cds.org

2 I did not use data from the 2019 edition because the relevance assessments were released only after I had finished this chapter; I believe that the additional data would not have changed any conclusions driven from this study though.

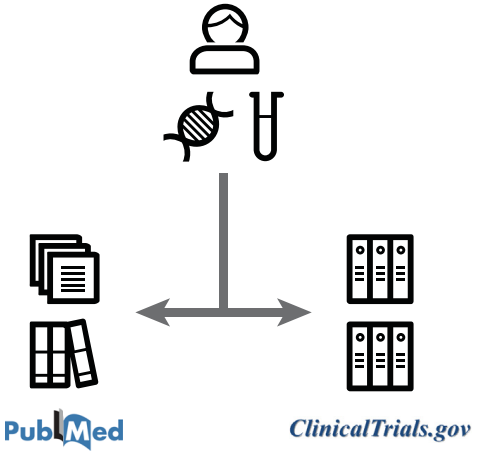

Figure 6.1: Overview of the TREC-PM task.

```
<topic number="38">
   <disease>cholangiocarcinoma</disease>
   <gene>IDH1</gene>
   <demographic>50-year-old male</demographic>
</topic>
```

Figure 6.2: A sample TREC-PM topic.

Each topic represented a patient with a given disease, gene, and demographic information[3]. For instance, Figure 6.2 depicts a 50-year-old man with cholangiocarcinoma (bile duct cancer) and a genotype associated with this condition, namely a mutation in the isocitrate dehydrogenase 1 (IDH1) gene.

Table 6.1 describes the number of documents and relevance assessments in the gold standard for each task edition. Relevance assessments were created by a group of medical students at the Oregon Health & Science University (OHSU) and postdoctoral fellows at the National Library of Medicine (NLM) [135]. Documents to be judged were selected using stratified random sampling with two strata: 100% of every document in the top-10 results and 15% (30% in 2017) of the documents in the ranks 11–55 (ranks 11–50 in 2017).

Due to the specialized nature of the task, results were judged in two tiers (see Figure 6.3). Firstly, assessors classified query-document pairs in different categories, namely (i) the document is about "precision medicine"; (ii) the disease is exactly, more generally, more specifically, or not at all mentioned in the document; (iii) the gene (and the gene variant) is exact, missing, or different; (iv) the demography matches, is excluded, or not discussed in the document. Secondly, a rule-based approach associated a relevance score to each query-document pair,

3 2017 topics included an additional *other* field with patient comorbidities. Since it was not used in later years, I decided not to consider it at all in this work to make experiments homogeneous across the two considered editions.

YEAR	2017	2018
Topics	30	50
Biomedical abstracts		
Documents	26 739 426	26 739 426
Sampling	100% in ranks 1–10	100% in ranks 1–10
	30% in ranks 11–50	15% in ranks 11–55
Relevance assessments	22 642	22 429
Definitely relevant	2022 (8.93%)	3442 (15.35%)
Partially relevant	1853 (8.18%)	2146 (9.57%)
Clinical trials		
Documents	241 006	241 006
Sampling	100% in ranks 1–15	100% in ranks 1–10
		15% in ranks 11–55
Relevance assessments	13 441	14 188
Definitely relevant	436 (3.24%)	873 (6.15%)
Partially relevant	735 (5.47%)	1174 (8.28%)

Table 6.1: TREC-PM data overview.

either *definitely relevant* (2), *partially relevant* (1), or *not relevant* (0). Definitely relevant documents must be about "precision medicine", mention an exact or more specific disease, an exact gene, and a matching or not discussed demography; partially relevant documents must also be about "precision medicine", but may have a more general disease and a missing or different gene variant; not relevant documents are everything else.

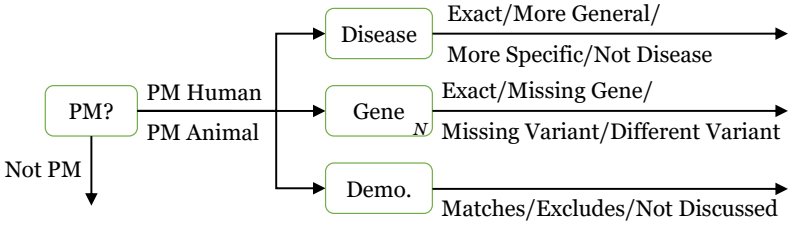

Figure 6.3: Overview of the TREC-PM assessment process. Reproduced from Roberts et al. [136] courtesy of the National Institute of Standards and Technology, U.S. Department of Commerce. Not copyrightable in the United States (public domain).

6.2.2 *Evaluation Metrics*

I assessed the experiments using the official evaluation script TREC_EVAL[4] and collected the metrics infNDCG (or NDCG for CT, since stratified sam-

4 https://github.com/usnistgov/trec_eval

pling was not available in 2017), precision at 10 results (P@10), and SET_RECALL — described below in detail (see also Section 3.2).

P@K P@K is defined as the precision at the top-k results, normally with $k = 10$.

CG Cumulative Gain (CG) is the sum of relevance assessments up to a position k.

DCG Discounted Cumulative Gain (DCG) discounts CG by taking the rank in which a document occurs into account and further penalizing relevant documents that appear lower in the search results.

NDCG Normalized Discounted Cumulative Gain (NDCG) normalizes DCG over different queries by the ratio to the ideal DCG, when documents are sorted.

INFNDCG inferred Normalized Discounted Cumulative Gain (infNDCG) uses stratified random sampling to estimate the NDCG of a search engine when only incomplete judgments are possible [143].

SET_RECALL SET_RECALL is the ratio of relevant documents retrieved in the result list.

While P@K and SET_RECALL use a binary notion of relevance (either a document is relevant to a given query or not), infNDCG and NDCG allow a graded relevance assessment, in which a given document can be more relevant than another for a given query. I used the official result list size of 1000 documents per query.

Precision-recall curve

I also plotted precision-recall curves, which are akin (but not equal) to Receiver Operating Characteristic (ROC) curves. Precision-recall curves plot precision versus recall at different points in the search results and thus summarizes the trade-off between recall and precision in the search engine. They are a preferred choice over ROC curves in the IR domain because the number of false negatives is normally very high [144]. They also provide here a more detailed picture over the search results than a single metric like R-PREC (the point where precision equals recall, visible in the plots) used in the official ranking.

6.3 METHODS

I explored three methods for query expansion, namely using word embeddings, rules, or terminologies and compared them to both a naive baseline and query boosting. I performed Fisher's randomization tests comparing the metrics across topics to check whether any differences were statistically significant, with $\alpha = 0.05$.

6.3.1 *Experimental Framework*

We built a Java framework in order to assess different ranking strategies. The framework allowed us to plug in several query expansion methods together with different boosting strategies (see Figure 6.4). Our source code is available at GitHub[5] under the MIT license.

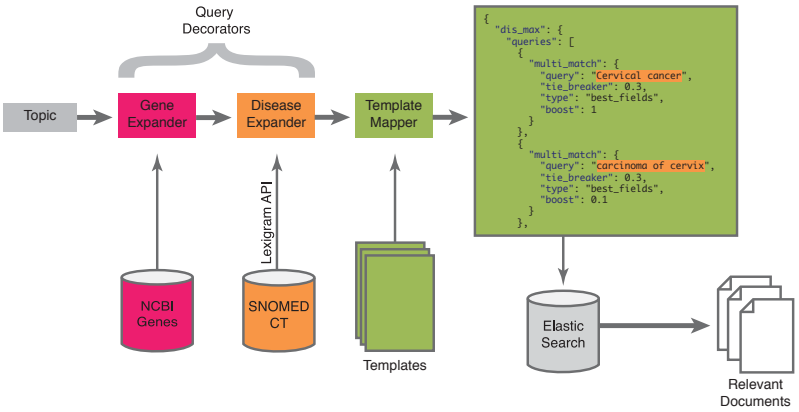

Figure 6.4: Framework structure implemented for TREC-PM experiments.

dis_max

We employed weighted `dis_max`[6] queries to improve recall without a corresponding drop in precision. In our framework, `dis_max` (see Equation 6.1) queries maximizes a disjunctive clause of weighted hypernyms and synonyms scores — namely t_o, original term; t_p, preferred term; s_i, synonyms; h_j, hypernyms; u, everything else — so that, e. g., the original term score receives more weight than one of its hypernyms ($\alpha \geqslant \beta \geqslant \gamma \geqslant \delta \gg \varepsilon$).

$$\text{dis\_max} = \max(\alpha t_o,\ \beta t_p,\ \gamma s_1,\ \ldots,\ \gamma s_n,\ \delta h_1,\ \ldots,\ \delta h_n,\ \varepsilon u) \quad (6.1)$$

6.3.2 *Elasticsearch*

We indexed data using ES[7] 5.4.0 and kept the default options wherever possible. ES is based on Apache Lucene[8], which pre-processes text with (1) lowercasing, (2) removal of a custom list of 33 stopwords, and (3) the Porter stemming algorithm (see Section 2.3.1).

5 https://github.com/JULIELab/trec-pm
6 https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-dis-max-query.html
7 https://www.elastic.co
8 https://lucene.apache.org/

Documents are ranked using the Okapi BM25 algorithm [145], a probabilistic variant of term frequency - inverse document frequency (tf-idf) (see Equation 6.2).

$$\mathrm{RSV}(q, d) = \sum_{t \in q} \log \frac{N - df_t + \frac{1}{2}}{df_t + \frac{1}{2}} \times \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b\frac{L_d}{L_{ave}}) + tf_{td}} \quad (6.2)$$

Here, the relevance RSV of a given document d is a function over all terms t of the query q, their document frequencies df_t, their term frequencies tf_{td}, and two hyperparameters, b and k_1 [13, Section 11.4.3]. In the equation, the first factor is an alternative version of inverse document frequency (idf) that measures how often a term occurs in the full collection of size N and penalizes common terms. Conversely, the second factor is an improved version of term frequency (tf) that takes into account the ratio of the document length L_d to the average length L_{ave} of a document in the collection. This ratio is controlled by the hyperparameter b ($0 \leqslant b \leqslant 1$): 0 disables length normalization, while 1 enables full scaling to the average length. Finally, k_1 controls the saturation of tf: 0 turns it into a binary model, while a large value allows tf to be used freely. Commonly, $b = 0.75$ and $k_1 = 1.2$.

6.3.3 *Baseline*

I set the baseline to a boolean `must` query[9] matching the topic disease and gene as-is and augmented the result list with any remaining document not previously matched. `Must` queries match a document only if the searched term appears in the document and thus provide a strict baseline.

6.3.4 *Query Boosting*

We used therapy and prognosis terms, as well as common suffixes of cancer drugs (e. g., **mab* → *monoclonal antibodies*) to boost relevant results (see Section 2.6). Conversely, we pushed down less relevant results with keywords related to in vitro research. The updated BM25 score $\mathrm{RSV}'(q, d)$ is calculated via Equation 6.3, where $\mathcal{P}$ is the set of positive keywords and $\mathcal{N}$ is the set of negative keywords.

$$\mathrm{RSV}'(q, d) = \mathrm{RSV}(q, d) + \sum_{p \in \mathcal{P}} \mathrm{RSV}(p, d) - \sum_{n \in \mathcal{N}} \mathrm{RSV}(n, d) \quad (6.3)$$

We first built a pool of candidate keywords using a combination of (a) medical knowledge, (b) manual result analysis, and (c) suffixes

9 https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-bool-query.html

of cancer drugs extracted from the code L section (L01 to L04) of the ATC[10] code list containing antineoplastic, immunomodulating, and experimental drugs. We then tested the keywords against the gold standard and selected only the ones that did not degrade any evaluation metric. Table 6.2 provides the final list of boosters used in the experiments.

POSITIVE BOOSTERS			NEGATIVE BOOSTERS	
surgery	resistance	-mab	transcript	probes
therapy	recurrence	-nib	paraffin	detection
treatment	targets	-cin	tumorigenesis	screening
prognosis	malignancy	-one	embedded	
prognostic	study	-ate	formalin	
survival	therapeutical	-mus	fish	
patient	outcome	-lin	tissue	

Table 6.2: List of positive and negative keyword boosters.

6.3.5 *Query Expansion*

Using Word Embeddings

I leveraged the proposed framework to enrich the disease and gene fields from the topics to improve recall (see Section 2.6). I experimented with expanding the disease and gene mentioned in each topic with the top three most-similar words according to a model based on word embeddings. I used the 200-dimensional word embeddings from BIO.NLPLAB.ORG (see Section 3.1.2), pre-trained on PUBMED data using WORD2VEC with a skip-gram model and the default hyperparameters.

Table 6.3 provides some examples of expansions obtained by querying the model with diseases and genes used in the TREC-PM challenge. It shows that it was able to correctly capture not only syntactic forms such as plural (e. g., "cholangiocarcinomas"), casing (e. g., "Glioblastoma"), acronyms (e. g., "GBM" for "glioblastoma multiforme"), and misspellings (e. g., "caner"), but also lexical variants such as synonyms (e. g., "LKB1" and "HER2" for the STK11 and ERBB2 genes, respectively) and hypernyms (e. g., "colorectal" for "colon" and the "BRCA" gene family for the BRCA1 gene).

Using Rules

I also deployed rule-generated variations of disease and gene, namely a regular expression to extract the gene family name and a rule to match names of solid tumors, to improve recall of clinical trials.

10 https://en.wikipedia.org/wiki/ATC_code_L

QUERY	MOST SIMILAR WORDS
cholangiocarcinoma	cholangiocarcinomas; hepatolithiasis; cholangiocellular; HCC; HCCa
colon cancer	colorectal; caner; cancers; adenocarcinoma; CRC
glioblastoma	glioma; medulloblastoma; GBM; Glioblastoma; glioblastomas
BRCA1	PALB2; BRCA2; 6174delT; BRCA; c.156_157insAlu
ERBB2	HER2; HER2/neu; Her-2/neu; erbB-2; HER-2/neu
STK11	STK11/LKB1; LKB1/STK11; HRPT2; TP63; FLCN

Table 6.3: Examples of query expansion using word embeddings provided by BIO.NLPLAB.ORG.

DISEASE NAME EXPANSION I noticed that some relevant clinical trials would mention only the top-level concept of "solid tumors" and not specify a unique disease. Given that the terminologies we used did not include such a mapping, I further expanded the disease term with "solid" following Equation 6.4, in which a topic is expanded if, and only if, it does not contain the words "lymphoma" or "leukemia".

$$\text{solid} \leftrightarrow \neg(\text{lymphoma} \vee \text{leukemia}) \tag{6.4}$$

GENE NAME EXPANSION Upon manual data inspection, I noticed that some clinical trials would target a whole gene family and not a single gene for the study (e. g., BRCA instead of BRCA1). Since terminologies containing this mapping are far from complete, I devised a regular expression to automatically perform a naive mapping to approximate the gene family. The pattern `([0-9]{1,2}[A-Z]{0,2}|R[0-9]{0,1})$` removes up to two trailing digits (optionally followed by up to two characters) or a trailing `R` character (optionally followed by a single digit). Table 6.4 shows some examples of gene families extracted.

GENE	FAMILY
BRCA2	BRCA
TP53	TP
CDK6	CDK
FGFR1	FGF
EGFR	EGF
PIK3CA	PIK
NF2	NF
CDKN2A	CDKN
EML4	EML

Table 6.4: Examples of gene family names extracted by the rule-based approach.

Using Terminologies

Finally, domain terminologies were leveraged for query expansion.

DISEASE NAME EXPANSION We performed disease name expansions with results provided by the LEXIGRAM[11] Application Programming Interface (API), largely based on the SNOMED CT, the MeSH, and the International Classification of Diseases (ICD). We expanded each disease term with its preferred term and synonyms.

GENE EXPANSION I performed gene name expansion using the description and synonyms columns of the National Center for Biotechnology Information (NCBI) Homo Sapiens Gene List[12]. This allowed, e. g., matches on documents mentioning the LKB1 gene for topics mentioning only the synonym STK11.

6.4 RESULTS

6.4.1 *Biomedical Abstracts*

Figure 6.5 shows boxplots across all topics of the different experiments performed with biomedical abstracts documents. A red square mark plots the average value, while dashed lines present the median and maximum values obtained over all participant runs, when available. When compared to the baseline, query boosting improved infNDCG (+0.0319, $p < 0.001$) and P@10 (+0.0925, $p < 0.001$) with a small improvement of SET_RECALL (+0.0030, $p > 0.05$). When compared to query boosting, query expansion methods slightly reduced infNDCG (embeddings: −0.0074, $p < 0.05$; rules: −0.0138, $p < 0.001$; terminologies: −0.0038, $p > 0.05$) and P@10 (embeddings: −0.0075, $p > 0.05$; rules: −0.0200, $p < 0.01$; terminologies: −0.0075, $p > 0.05$), but increased recall (embeddings: +0.0150, $p > 0.05$; rules: +0.0306, $p < 0.05$; terminologies: −0.0445, $p < 0.001$). Terminologies provided the largest improvement in SET_RECALL with the smallest reduction in infNDCG and P@10. Rules provided the second largest improvement in SET_RECALL, but affected infNDCG and P@10 the most.

Figure 6.6 displays the precision-recall curves for the experiments mentioned before. It shows that all methods improved over the baseline mostly in the top results, as demonstrated before by the large improvement in P@10.

Finally, Figure A.1 (see Appendix A) depicts the metrics of the proposed experiments detailed by topic and compared to the median and best values obtained by participant runs, shown in dashed red lines when available. Considering infNDCG, query boosting improved

11 https://www.lexigram.io

12 ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

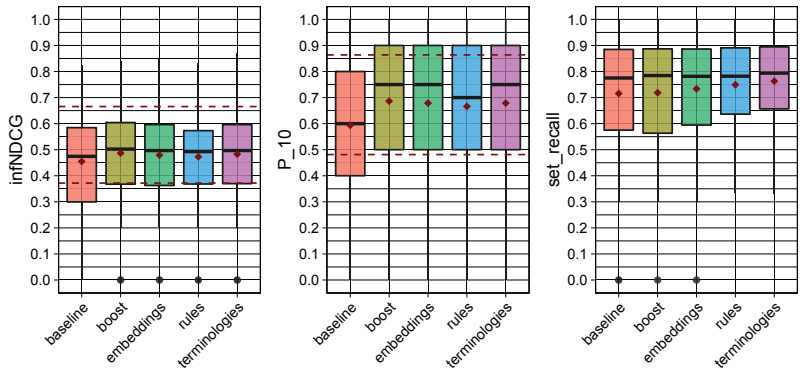

Figure 6.5: Biomedical abstracts: distribution of infNDCG, P@10, and recall across 80 topics.

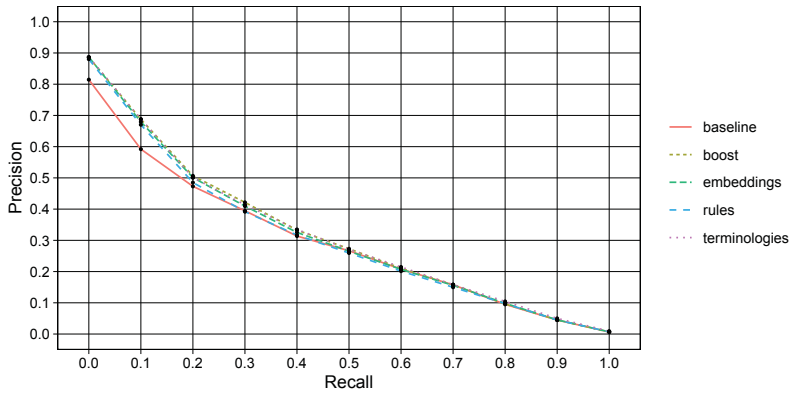

Figure 6.6: Biomedical abstracts: averaged eleven-point precision-recall curves across 80 topics.

topic 201717 (prostate cancer, PTEN inactivating) the most (+0.2350), while it worsened topic 201825 (melanoma, high serum LDH levels) the most (−0.1226). Meanwhile, compared to query boosting, embeddings improved topic 201719 (Colorectal cancer, FGFR1 Amplification) the most (+0.0624) and topic 201835 (breast cancer, CDKN2A) the least (−0.1414); rules improved topic 201715 (Cervical cancer, STK11) the most (+0.1725) and topic 201835 the least (−0.1615); terminologies improved topic 201715 (Cervical cancer, STK11) the most (+0.1515) and topic 201835 (breast cancer, CDKN2A) the least (-0.1613). Compared to official runs, the methods presented here performed overall worse (still concerning infNDCG) in the topics from 201801 to 201825, a subset of topics about melanoma.

Considering SET_RECALL, topics 201836 (lung cancer, ERBB2), 201715 (Cervical cancer, STK11), and 201831 (head and neck squamous cell carcinoma, CDKN2A) observed the largest improvements for rules (+0.3549, +0.4000, and +0.3810, respectively) and terminologies (+0.4516, +0.4000, and +0.3810, respectively). Embeddings helped topics 201836 (lung cancer, ERBB2), 201722 (Lung cancer, ERBB2 Amplification), and 201843 (basal cell carcinoma, PTCH1) the most (+0.4194, +0.3802, and +0.2800, respectively). This is probably explained by the prevalence of common gene synonyms, namely HER2 for ERBB2 and LKB1 for STK11.

6.4.2 *Clinical Trials*

Similar to biomedical abstracts, Figure 6.7 shows boxplots across all topics of the different experiments performed with clinical trials. When compared to the baseline, query boosting improved NDCG (+0.0094, $p > 0.05$) and SET_RECALL (+0.0536, $p > 0.05$) with a small worsening of P@10 (−0.0163, $p > 0.05$). When compared to query boosting, query expansion methods improved NDCG (embeddings: +0.0094, $p > 0.05$; rules: +0.0312, $p < 0.05$; terminologies: +0.0238, $p < 0.05$) and SET_RECALL (embeddings: +0.0536, $p < 0.01$; rules: +0.1075, $p < 0.001$; terminologies: +0.0843, $p < 0.001$), but slightly reduced P@10 (embeddings: −0.0163, $p > 0.05$; rules: −0.0126, $p > 0.05$; terminologies: −0.0151, $p > 0.05$). Rules provided the largest improvement in SET_RECALL and NDCG, with the smallest reduction in P@10. Terminologies provided the second largest improvement in SET_RECALL and NDCG. Embeddings affected P@10 the most.

Figure 6.8 displays the precision-recall curves for the experiments mentioned before. It shows that (1) all methods performed slightly worse than both the baseline and query boosting; and (2) query boosting was better than the baseline only in the initial results.

Finally, Figure A.2 (see Appendix A) depicts the metrics of the proposed experiments detailed by topic and compared to the median and best values obtained by participant runs, shown in dashed red

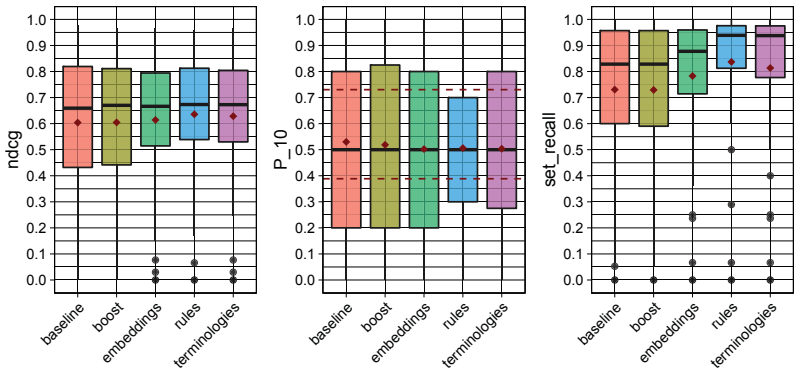

Figure 6.7: Clinical trials: distribution of NDCG, P@10, and recall across 80 topics.

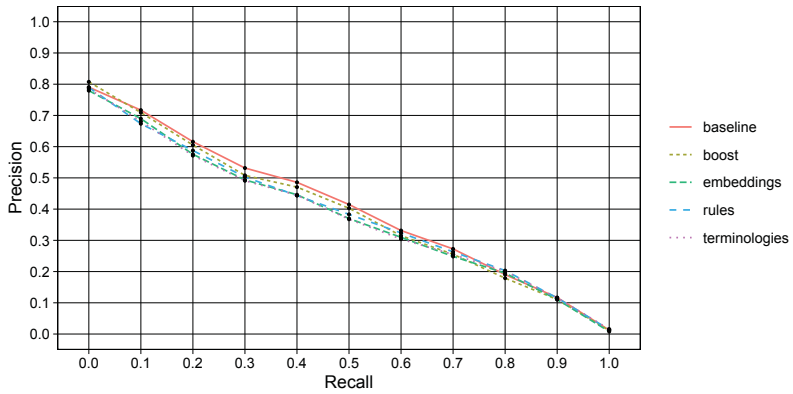

Figure 6.8: Clinical trials: averaged eleven-point precision-recall curves across 80 topics.

lines when available. Considering NDCG, query boosting improved topic 201723 (Breast cancer, PTEN loss) the most (+0.1494), while it worsened topic 201804 (melanoma, BRAF (K601E)) the most (−0.0634). Meanwhile, compared to query boosting, embeddings improved topic 201722 (Lung cancer, ERBB2 Amplification) the most (+0.3768) and topic 201835 (breast cancer, CDKN2A) the least (−0.2872); rules improved topic 201842 (glioblastoma, CDK6) the most (+0.4105) and topic 201835 the least (−0.3563); terminologies improved topic 201722 (Lung cancer, ERBB2 Amplification) the most (+0.3810) and topic 201835 (breast cancer, CDKN2A) the least (−0.3314).

Considering SET_RECALL, topics 201836 (lung cancer, ERBB2), 201722 (Lung cancer, ERBB2 Amplification), and 201728 (Pancreatic ductal adenocarcinoma, ERBB3) observed the largest improvements for embeddings (+0.5405, +0.5196, +0.5000, respectively), rules (+0.6486, +0.5882, and +0.5000, respectively), and terminologies (+0.5405, +0.5392, and +0.5000, respectively).

6.5 DISCUSSION

Biomedical abstracts and clinical trials seem to behave in opposite ways. While query boosting increases precision for the former, the same is not seen for the latter; conversely, query expansion increases recall for clinical trials much more prominently than for biomedical abstracts. This is probably explained by the huge difference in collection size[13] and number of relevant documents (given a particular query): PubMed is a large collection with a sizable number of both relevant and irrelevant documents, while ClinicalTrials.gov is a smaller data set with just a few (if any) relevant documents. Query boosting thus helped in prioritizing biomedical abstracts, while query expansion supported the retrieval of the few relevant trials. Retrieving these trials then improves overall NDCG as a side effect because most of the time there are no other possible relevant documents. Nonetheless, the common usage of `dis_max` queries allowed us to successfully perform query expansion without a drop in precision, even in the biomedical abstracts subtask.

Concerning the different query expansion strategies and considering overall retrieval quality, the hand-crafted rules (which provided only hypernyms) benefited clinical trials more strongly than biomedical abstracts, while word embeddings and terminologies (which provide both synonyms and hypernyms) were more prone to help to retrieve biomedical abstracts. This could once again be justified by the diverging goals of the two collections: on the one hand, PubMed abstracts tend to be very specific by describing a narrow clinical condition (e. g., the involvement of BRCA2 in prostate cancer); on the other hand,

13 Similar findings were also reported by Ghawi and Pfeffer [146] and Faessler et al. [147].

clinical trials try to broaden the scope to enroll enough patients (e. g., patients with any BRCA gene family mutation and any solid tumor).

Overall, word embeddings could be an easy shortcut for improved recall in the absence of formal domain thesauri or interest for rule maintenance. They could furthermore be used to bootstrap a terminology and assist the creation of human-curated resources. Lastly, they may improve the quality of results obtained from small collections while (with proper query engineering) prevent worsening it in larger databases.

6.5.1 *Error Analysis*

I investigated in deeper detail the documents retrieved for three topics that consistently showed lower metrics with query expansion when compared to the query boosting baseline.

BREAST CANCER, CDKN2A Every query expansion method decreased all metrics for topic 201835 for both biomedical abstracts and clinical trials. Even though word embeddings correctly captured the synonym CDKN2 for the gene (also provided by the NCBI gene list), they also mapped to two other well-known tumor suppressor genes (RB1 and TP53) that fail to capture the specificity required for this task. Precision was further deteriorated by disease embeddings expanding breast cancer to prostate and colorectal cancer, also not precise enough for this task. Moreover, I later identified that "CDK" would be a better family for the gene in the rule-based approach, thus allowing a better ranking of definitely relevant documents such as `NCT03050398`[14]. Last, I required exact matches on synonyms from terminologies (to avoid a reduction in precision), but this prevented partial matches on expansions such as "cyclin-dependent kinase 4 inhibitor A". The problem was further aggravated by the low number of relevant documents present in the gold standard (29 for biomedical abstracts and 4 for clinical trials), which dramatically increases the odds of incorrect addition of irrelevant documents with query expansion methods.

NON-SMALL CELL CARCINOMA, MET Metrics for topic 201837 were also penalized by every query expansion method for both subtasks. Once again, gene embeddings correctly captured the synonym "c-MET", but reduced precision by providing expansions to two other receptors of tyrosine kinases, namely EGFR and HER1, but surprisingly not HGFR (a proper synonym provided by NCBI). Rules were not able to capture the gene family given the nonexistence of digits in the gene; LEXIGRAM did not provide any expansions for the disease. Better expansion candidates might have been provided if the topic correctly mentioned a "non-small cell lung carcinoma", the most com-

14 https://clinicaltrials.gov/ct2/show/NCT03050398

mon disorder associated with this morphology. Finally, I also rejected the hypothesis that the worse metrics owed to a reduced number of relevant documents in the gold standard.

PAPILLARY THYROID CARCINOMA, NTRK1 Topic 201841 showed an anomalous behavior: while all query expansion methods improved set recall for clinical trials (and embeddings for biomedical abstracts), other metrics were negatively affected. Considering the gene, only the NCBI gene list was able to capture the TRK family, while embeddings provided expansions to different genes such as NTRK3, RET, and its common mutation C620S in the transmembrane domain, all associated with papillary thyroid carcinoma. Similar to topic 201835, this topic had also a low number of relevant documents in the gold standard, viz. six biomedical abstracts and 14 clinical trials.

6.5.2 *Shared Task Results*

Compared to all 32 participating teams in 2017 [135] (27 in 2018 [136] and 15 in 2019 [148]) and considering the ranks on the three evaluated metrics, we had in the first year the second-best average rank (2.7); in the second year, the best average rank (1.7) and the best P@10 in the biomedical abstracts subtask (see Table 6.5); and, in the third year, the best metrics overall. Conversely, in the clinical trials subtask (see Table 6.6), we had in 2018[15] the third-best average rank (4.0) with best infNDCG and in 2019, except for P@10 (in which we ranked second), the best metrics overall.

Note that the results presented here differ from the submitted results in different ways, such as:

- We did not employ `dis_max` queries in 2017;
- We performed query expansion using terminologies associated with rules;
- We fine-tuned weights and matching type (exact, lenient) for the submission using an internal gold standard;
- Our team explored in 2018 a supervised classifier for "precision medicine" and included it in two runs.

6.5.3 *Limitations and Future Work*

Even though negation is commonly found in clinical trials, we addressed it only partially by discarding exclusion criteria clearly marked in the data set. However, a closer inspection of the documents provided several examples where exclusion statements were also mentioned

15 We did not participate in the clinical trials subtask in 2017.

YEAR	TEAM	INFNDCG	R-PREC	P@10	AVERAGE RANK
2019	**julie-mug**	**0.5783 (1)**	**0.3572 (1)**	**0.6525 (1)**	**1.0**
	BITEM_PM	0.5339 (2)	0.3166 (3)	0.6275 (3)	2.7
	DUTIR	0.5108 (4)	0.3273 (2)	0.5975 (4)	3.3
	CCNL	0.5309 (3)	0.3066 (7)	0.6500 (2)	4.0
	imi_mug	0.4812 (5)	0.3122 (4)	0.5750 (6)	5.0
2018	**hpi-dhc**	**0.5605 (2)**	**0.3658 (2)**	**0.7060 (1)**	**1.7**
	MedIER	0.5515 (4)	0.3684 (1)	0.6220 (5)	3.3
	Cat_Garfield	0.5621 (1)	0.3257 (7)	0.6680 (2)	3.3
	SIBTextMining	0.5410 (5)	0.3574 (5)	0.6320 (3)	4.3
	UCAS	0.5580 (3)	0.3654 (3)	0.5980 (8)	4.7
2017	UTDHLTRI	0.4647 (1)	0.2993 (1)	0.6300 (2)	1.3
	imi_mug	**0.4158 (3)**	**0.2772 (2)**	**0.6267 (3)**	**2.7**
	BiTeM	0.4175 (2)	0.2687 (3)	0.5500 (4)	3.0
	UD_GU_BioTM	0.4135 (4)	0.2477 (7)	0.6400 (1)	4.0
	prna-mit-suny	0.4070 (5)	0.2622 (4)	0.5300 (5)	4.7

Table 6.5: Biomedical abstracts: top overall systems in the three editions of TREC-PM.

YEAR	TEAM	INFNDCG	R-PREC	P@10	AVERAGE RANK
2019	**julie-mug**	**0.6451 (1)**	**0.4820 (1)**	**0.5474 (2)**	**1.3**
	ims_unipd	0.6239 (2)	0.4386 (2)	0.5368 (3)	2.3
	cbnu	0.5568 (3)	0.4121 (3)	0.4921 (6)	3.7
	ECNU-ICA	0.5355 (4)	0.4001 (4)	0.5053 (4)	4.0
	BITEM_PM	0.4963 (6)	0.3698 (5)	0.4711 (7)	6.0
2018	Cat_Garfield	0.5503 (2)	0.4294 (1)	0.6260 (1)	1.3
	ims_unipd	0.5395 (3)	0.4128 (2)	0.5660 (2)	2.3
	hpi-dhc	**0.5545 (1)**	**0.4081 (4)**	**0.5400 (7)**	**4.0**
	Poznan	0.4894 (7)	0.4101 (3)	0.5580 (3)	4.3
	UCAS	0.5347 (4)	0.4005 (5)	0.5460 (6)	5.0
		P@5	P@10	P@15	
2017	UD_GU_BioTM	0.5448 (1)	0.4448 (1)	0.3885 (1)	1.0
	UTDHLTRI	0.4483 (4)	0.4172 (2)	0.3816 (2)	2.7
	teckro	0.4276 (8)	0.4000 (3)	0.3632 (3)	4.7
	udel	0.4552 (2)	0.3793 (5)	0.3333 (8)	5.0
	NOVASearch	0.4414 (6)	0.3966 (4)	0.3448 (5)	5.0

Table 6.6: Clinical trials: top overall systems in the three editions of TREC-PM.

in the inclusion criteria, as well as documents in which both criteria were mentioned in a single section. We believe that detecting negated sentences with NegEx or LingScope could improve overall metrics.

Results can also be easily affected by the quality of expansions provided for melanoma alone. This is because 27 out of the 80 topics considered in the two TREC-PM editions mention this condition, while in reality breast and prostate cancers are much more common. A more robust approach would be to group results per cancer topography or normalize by, e. g., overall incidence.

I also did not consider the impact of a different number of word embeddings nor randomization effects that may change the set produced. Three expansions not only are suggested in literature but in practice also closely resembles the practical amount of expansions found with terminologies: on average, we found 2.54 synonyms for diseases using Lexigram and 4.92 gene symbols using the NCBI gene list. Nonetheless, a stronger strategy would be to train several models and pool a common subset among them.

Future work should also consider the effect of local query expansion by using, e. g., relevance feedback. That would require, however, using the gold standard to provide annotations, thus turning it into a supervised approach. Alternatively, one could explore an automatic variant thereof, namely pseudo relevance feedback, which assumes that the top-k documents retrieved are relevant.

I would like to also explore the impact of the Okapi BM25 hyperparameters b and k_1 on clinical information retrieval. I hypothesize that length normalization may not play a role as important as in the general domain, since abstracts and trials share a similar length among themselves and, when not, it does not signal quality. This experiment, although simple, would require significant computational resources, since data would need to be indexed for every parameter combination.

7 CONCLUSION

7.1 SYNOPSIS

In this thesis, I have exploited the crucial role human language plays in healthcare and biomedical research documentation to address the needs of data-intensive precision medicine. More specifically, I have explored word embeddings to improve three main Natural Language Processing (NLP) tasks in this domain, namely clinical text cleansing, clinical text classification, and biomedical information retrieval.

In Chapter 4, I demonstrated a novel method using word embeddings to expand acronyms in an unsupervised way without relying on sense inventories. Furthermore, I proposed a minimal set of filtering rules to improve the precision of expansion candidates. This combined method outperformed traditional approaches that use either n-grams or a hand-crafted sense inventory.

In Chapter 5, I studied several methods for classification of longitudinal medical records. In particular, I verified that logistic regression associated with word embeddings constituted a better model for clinical text classification than more complex Deep Learning (DL) architectures. Moreover, I showed that word embeddings pre-trained on a larger corpus were not better than embeddings trained on the target dataset.

In Chapter 6, we proposed a method for increased recall that does not reduce the precision of results in a biomedical information retrieval scenario. Using this method, I demonstrated that word embeddings could be effectively used for query expansion when terminological resources were not available. I also revealed that the benefit of query expansion was stronger in a small dataset.

7.2 FINDINGS

ON DATA REQUIREMENTS Contrary to expected, the training of word embeddings does not require a huge corpus and can effectively be done on datasets with as few as millions of tokens. I could train not only embeddings that capture the meaning of acronyms using a collection of less than 30 000 documents with around eight million tokens (Chapter 4) but also embeddings used as features for logistic regression on a training set with 200 patient notes and less than 800 000 tokens (Chapter 5). Furthermore, I showed that embeddings trained on much larger corpora did not improve the efficiency of a downstream task (Chapter 5).

ON ANNOTATED DATA Since word embeddings leverage latent knowledge present in texts in an unsupervised fashion, training data can more easily be obtained without the need for large annotation efforts by human experts. With zero labels, embeddings successfully captured the meaning of acronyms (Chapter 4), as well as of diseases and genes (Chapter 6); when used as features in a downstream task, they effectively compressed all input data with a smaller loss (Chapter 5). Nonetheless, embeddings pre-trained on a large collection can easily be reused, which not only saves computational power but also allows transferring general knowledge to a small set.

ON PREPROCESSING Word embeddings supersede traditional preprocessing methods such as stemming/lemmatization, spell correction, short form resolution, and concept normalization by looking at the context in which variants occur and clustering them nearby in the vector space. I showed that they correctly map German inflections, misspellings, and clinical spelling variants to the same region in space (Chapter 4) and supported the claim that they also capture disease (Chapter 4 and Chapter 6) and gene synonymy (Chapter 6). Finally, I intrinsically evaluated their accuracy when expanding acronyms and thus endorsed their further reuse as features in downstream tasks (Chapter 4).

ON OUT-OF-VOCABULARY TERMS When trained with subword information, embeddings can exploit word morphology to provide an approximate representation even for words not previously seen in the training set, thus avoiding out-of-vocabulary issues due to composition or misspellings. I leveraged this property to overcome a small dataset in Chapter 5 and noted the clustering of additional rarer misspellings, which might be an undesired effect in some tasks such as acronym expansion (Chapter 4).

ON BASELINE APPROACHES Word embeddings constitute a new baseline approach for clinical NLP. When compared to frequency-based approaches, they overcame term frequency (tf) for acronym expansion (Chapter 4) and were not worse than term frequency - inverse document frequency (tf-idf) for text classification (Chapter 5). Since they compress additional contextual information in a vector space with fixed dimensions, I support their upfront use for textual feature representation.

7.3 OUTLOOK

Future work should consider more recent transfer learning methods, such as BERT [149], RoBERTa [150], and ELECTRA [151]. BERT and its variants are based on the Transformer model [152], a sequence

model that leverages attention to learn long-term dependencies. BERT not only provides contextualized word embeddings (e. g., specific embeddings for each sense of "cold") but also allows fine-tuning pre-trained models on a specific task. For instance, researchers could benefit from specific models for the biomedical domain using either BioBERT [153], ClinicalBERT [154, 155], or BlueBERT [30], pre-trained on all PubMed and MIMIC-III data. BERT thus enables more complex models to be applied to the typically small datasets from the clinical domain.

When dealing with languages other than English, one can either use multilingual models provided by, e. g., the original BERT authors or language-specific models such as GermanBERT[1] and CamemBERT [156]. In the biomedical domain, these models can then be further fine-tuned, e. g., on a language-specific subset of PubMed or directly on the target dataset.

Likewise, there is an increasing demand for understanding causality and creating explainable models in the clinical domain. When rule-based models are not a viable option, shallow Machine Learning (ML) models may provide a simpler decision boundary than DL strategies, with a small (if any) difference in the efficiency. On intrinsically complex problems, ablation studies may help researchers produce distilled models or locally interpretable explanations [157]. Overall, such approaches may help elucidate why DL underperforms for text classification and how language, data size, and data quality influence its behavior.

Given that the clinical domain is especially rich in terminological resources, another opportunity for future research lies on sense and concept embeddings [158]. For instance, the MeSH vocabulary could be used to build or enhance word embeddings with senses, while SNOMED CT could be exploited to create concept embeddings that leverage their attributes and semantic relations. Alternatively, joint models such as ERNIE [159] may be trained on text automatically enriched with entities to map words and concepts to the same vector space. Such semantic knowledge may aid in dealing with linguistic phenomena such as antonymy by providing clues on the relation among terms, thus bringing the NLP and knowledge representation research communities even closer.

1 https://deepset.ai/german-bert

A

APPENDIX

A.1 CLINICAL TEXT CLASSIFICATION

Table A.1 to Table A.7 present detailed results of the National NLP Clinical Challenges (n2c2) by target class for the seven explored strategies.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.0000	0.0000	0.0000	0.6512	1.0000	0.7887	0.3944	0.6512
ADVANCED-CAD	0.5233	1.0000	0.6870	0.0000	0.0000	0.0000	0.3435	0.5233
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ASP-FOR-MI	0.7907	1.0000	0.8831	0.0000	0.0000	0.0000	0.4416	0.7907
CREATININE	0.0000	0.0000	0.0000	0.7209	1.0000	0.8378	0.4189	0.7209
DIETSUPP-2MOS	0.5116	1.0000	0.6769	0.0000	0.0000	0.0000	0.3385	0.5116
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ENGLISH	0.8488	1.0000	0.9182	0.0000	0.0000	0.0000	0.4591	0.8488
HBA1C	0.0000	0.0000	0.0000	0.5930	1.0000	0.7445	0.3723	0.5930
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.5000	1.0000	0.6667	0.0000	0.0000	0.0000	0.3333	0.5000
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	0.0000	0.0000	0.0000	0.9070	1.0000	0.9512	0.4756	0.9070
Overall (macro)	0.3184	0.4615	0.3703	0.4463	0.5385	0.4836	0.4270	0.7648
Overall (micro)	0.6899	0.7756	0.7303	0.8289	0.7572	0.7914	0.7608	0.7648

Table A.1: Official results: Baseline.

A.2 BIOMEDICAL INFORMATION RETRIEVAL

Figure A.1 and Figure A.2 present detailed results of the TREC-PM challenge by topic for the Biomedical Abstract (BA) and Clinical Trial (CT) subtasks, respectively.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.8333	0.8333	0.8333	0.9107	0.9107	0.9107	0.8720	0.8837
ADVANCED-CAD	0.8000	0.8000	0.8000	0.7805	0.7805	0.7805	0.7902	0.7907
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9647	0.9880	0.9762	0.4881	0.9535
ASP-FOR-MI	0.8500	1.0000	0.9189	1.0000	0.3333	0.5000	0.7095	0.8605
CREATININE	0.6786	0.7917	0.7308	0.9138	0.8548	0.8833	0.8071	0.8372
DIETSUPP-2MOS	0.9111	0.9318	0.9213	0.9268	0.9048	0.9157	0.9185	0.9186
DRUG-ABUSE	0.5000	0.3333	0.4000	0.9762	0.9880	0.9820	0.6910	0.9651
ENGLISH	0.9359	1.0000	0.9669	1.0000	0.6154	0.7619	0.8644	0.9419
HBA1C	1.0000	0.8571	0.9231	0.9107	1.0000	0.9533	0.9382	0.9419
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.8085	0.8837	0.8444	0.8718	0.7907	0.8293	0.8369	0.8372
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	1.0000	0.6250	0.7692	0.9630	1.0000	0.9811	0.8752	0.9651
Overall (macro)	0.7140	0.6966	0.6993	0.8629	0.7820	0.8057	0.7525	0.9123
Overall (micro)	0.8784	0.9129	0.8953	0.9376	0.9120	0.9246	0.9100	0.9123

Table A.2: Official results: Rule-based classifier.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.5000	0.4333	0.4643	0.7167	0.7679	0.7414	0.6028	0.6512
ADVANCED-CAD	0.7115	0.8222	0.7629	0.7647	0.6341	0.6933	0.7281	0.7326
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ASP-FOR-MI	0.8310	0.8676	0.8489	0.4000	0.3333	0.3636	0.6063	0.7558
CREATININE	0.5000	0.5000	0.5000	0.8065	0.8065	0.8065	0.6532	0.7209
DIETSUPP-2MOS	0.5952	0.5682	0.5814	0.5682	0.5952	0.5814	0.5814	0.5814
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ENGLISH	0.8488	1.0000	0.9182	0.0000	0.0000	0.0000	0.4591	0.8488
HBA1C	0.5862	0.4857	0.5312	0.6842	0.7647	0.7222	0.6267	0.6512
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.7391	0.7907	0.7640	0.7750	0.7209	0.7470	0.7555	0.7558
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	1.0000	0.2500	0.4000	0.9286	1.0000	0.9630	0.6815	0.9302
Overall (macro)	0.5598	0.5168	0.5195	0.6595	0.6633	0.6602	0.5899	0.8095
Overall (micro)	0.7651	0.7734	0.7692	0.8410	0.8346	0.8378	0.8035	0.8095

Table A.3: Official results (test set): Support Vector Machine (SVM).

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.4583	0.3667	0.4074	0.6935	0.7679	0.7288	0.5681	0.6279
ADVANCED-CAD	0.6842	0.8667	0.7647	0.7931	0.5610	0.6571	0.7109	0.7209
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ASP-FOR-MI	0.8243	0.8971	0.8592	0.4167	0.2778	0.3333	0.5962	0.7674
CREATININE	0.5769	0.6250	0.6000	0.8500	0.8226	0.8361	0.7180	0.7674
DIETSUPP-2MOS	0.6170	0.6591	0.6374	0.6154	0.5714	0.5926	0.6150	0.6163
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ENGLISH	0.8488	1.0000	0.9182	0.0000	0.0000	0.0000	0.4591	0.8488
HBA1C	0.4800	0.3429	0.4000	0.6230	0.7451	0.6786	0.5393	0.5814
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.7037	0.8837	0.7835	0.8438	0.6279	0.7200	0.7518	0.7558
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	0.0000	0.0000	0.0000	0.9070	1.0000	0.9512	0.4756	0.9070
Overall (macro)	0.4737	0.5109	0.4887	0.6671	0.6441	0.6509	0.5698	0.8068
Overall (micro)	0.7537	0.7865	0.7697	0.8466	0.8209	0.8336	0.8017	0.8068

Table A.4: Official results: SELF-LR.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.5238	0.3667	0.4314	0.7077	0.8214	0.7603	0.5959	0.6628
ADVANCED-CAD	0.6610	0.8667	0.7500	0.7778	0.5122	0.6176	0.6838	0.6977
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9651
ASP-FOR-MI	0.8267	0.9118	0.8671	0.4545	0.2778	0.3448	0.6060	0.7791
CREATININE	0.6250	0.6250	0.6250	0.8548	0.8548	0.8548	0.7399	0.7907
DIETSUPP-2MOS	0.6250	0.6818	0.6522	0.6316	0.5714	0.6000	0.6261	0.6279
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9647	0.9880	0.9762	0.4881	0.9535
ENGLISH	0.8488	1.0000	0.9182	0.0000	0.0000	0.0000	0.4591	0.8488
HBA1C	0.5172	0.4286	0.4688	0.6491	0.7255	0.6852	0.5770	0.6047
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.7059	0.8372	0.7660	0.8000	0.6512	0.7179	0.7420	0.7442
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	0.0000	0.0000	0.0000	0.9070	1.0000	0.9512	0.4756	0.9070
Overall (macro)	0.4845	0.5167	0.4970	0.6702	0.6463	0.6531	0.5751	0.8113
Overall (micro)	0.7583	0.7930	0.7753	0.8511	0.8240	0.8373	0.8063	0.8113

Table A.5: Official results: PRE-LR.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.3429	0.4000	0.3692	0.6471	0.5893	0.6168	0.4930	0.5233
ADVANCED-CAD	0.5932	0.7778	0.6731	0.6296	0.4146	0.5000	0.5865	0.5465
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9651	1.0000	0.9822	0.4911	0.9535
ASP-FOR-MI	0.7922	0.8971	0.8414	0.2222	0.1111	0.1481	0.4948	0.7442
CREATININE	0.2581	0.3333	0.2909	0.7091	0.6290	0.6667	0.4788	0.5698
DIETSUPP-2MOS	0.5918	0.6591	0.6237	0.5946	0.5238	0.5570	0.5903	0.6047
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9643	0.9759	0.9701	0.4850	0.9651
ENGLISH	0.8571	0.9863	0.9172	0.5000	0.0769	0.1333	0.5253	0.8372
HBA1C	0.3889	0.2000	0.2642	0.5882	0.7843	0.6723	0.4682	0.6047
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.5000	0.6512	0.5657	0.5000	0.3488	0.4110	0.4883	0.5349
MAKES-DECISIONS	0.9651	1.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.9651
MI-6MOS	0.0000	0.0000	0.0000	0.9036	0.9615	0.9317	0.4658	0.9070
Overall (macro)	0.4069	0.4542	0.4252	0.6326	0.5704	0.5838	0.5045	0.7504
Overall (micro)	0.6700	0.7298	0.6986	0.7994	0.7496	0.7737	0.7362	0.7504

Table A.6: Official results: SELF-LSTM.

	MET			NOT MET			OVERALL	
CRITERION	P	R	F_1	P	R	F_1	F_1	A
ABDOMINAL	0.3704	0.3333	0.3509	0.6610	0.6964	0.6783	0.5146	0.6047
ADVANCED-CAD	0.5417	0.8667	0.6667	0.5714	0.1951	0.2909	0.4788	0.5465
ALCOHOL-ABUSE	0.0000	0.0000	0.0000	0.9647	0.9880	0.9762	0.4881	0.9651
ASP-FOR-MI	0.7907	1.0000	0.8831	0.0000	0.0000	0.0000	0.4416	0.7791
CREATININE	0.3226	0.4167	0.3636	0.7455	0.6613	0.7009	0.5322	0.6395
DIETSUPP-2MOS	0.4783	0.5000	0.4889	0.4500	0.4286	0.4390	0.4640	0.4651
DRUG-ABUSE	0.0000	0.0000	0.0000	0.9647	0.9880	0.9762	0.4881	0.9651
ENGLISH	0.8554	0.9726	0.9103	0.3333	0.0769	0.1250	0.5176	0.8488
HBA1C	0.4400	0.3143	0.3667	0.6066	0.7255	0.6607	0.5137	0.5465
KETO-1YR	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	0.5000	1.0000
MAJOR-DIABETES	0.5400	0.6279	0.5806	0.5556	0.4651	0.5063	0.5435	0.5465
MAKES-DECISIONS	0.9647	0.9880	0.9762	0.0000	0.0000	0.0000	0.4881	0.9651
MI-6MOS	0.0000	0.0000	0.0000	0.9048	0.9744	0.9383	0.4691	0.9070
Overall (macro)	0.4080	0.4630	0.4298	0.5967	0.5538	0.5609	0.4953	0.7522
Overall (micro)	0.6680	0.7407	0.7025	0.8046	0.7436	0.7729	0.7377	0.7522

Table A.7: Official results: PRE-LSTM.

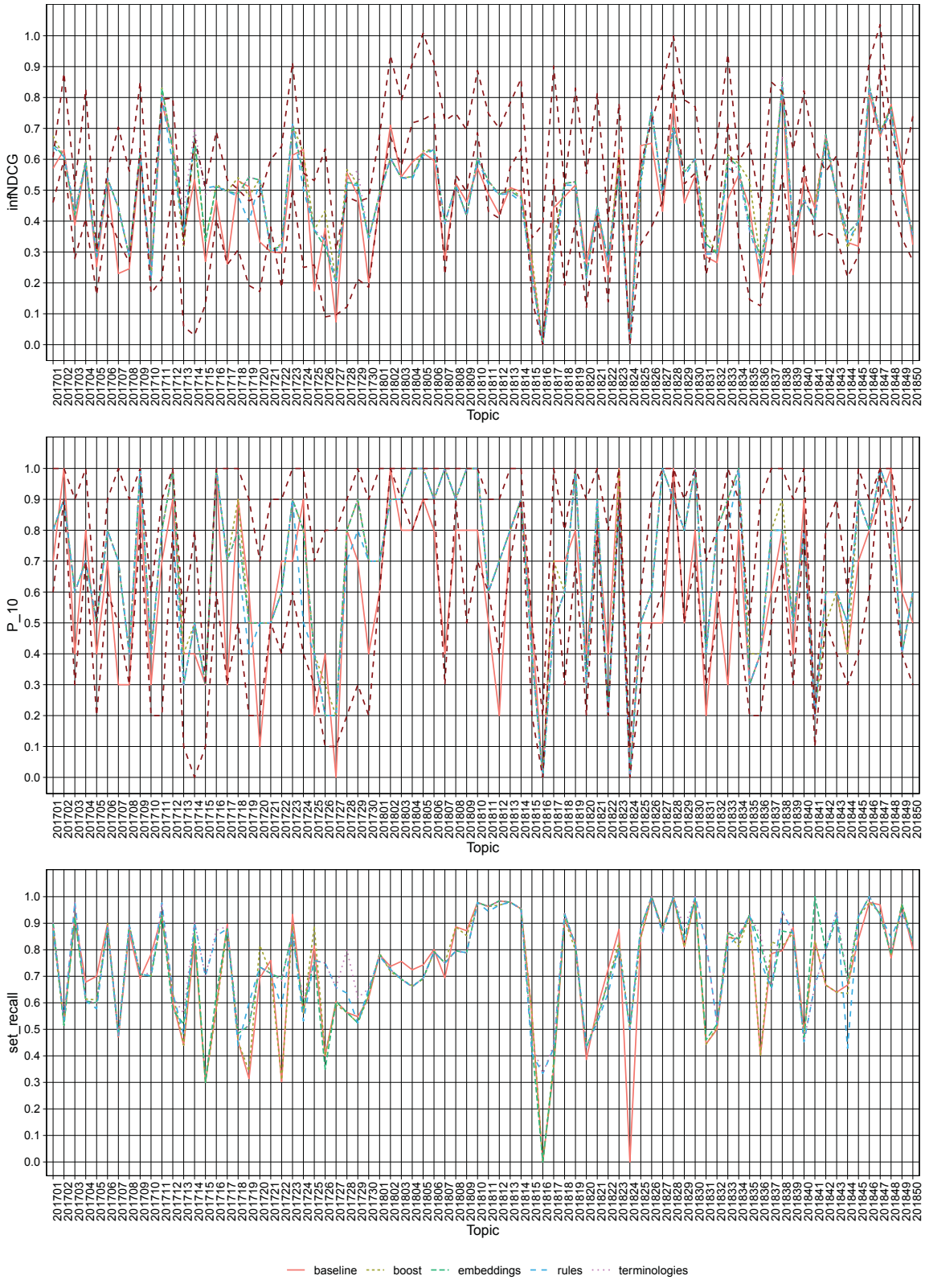

Figure A.1: Biomedical abstracts: Normalized Discounted Cumulative Gain (NDCG), P@10, and recall per topic.

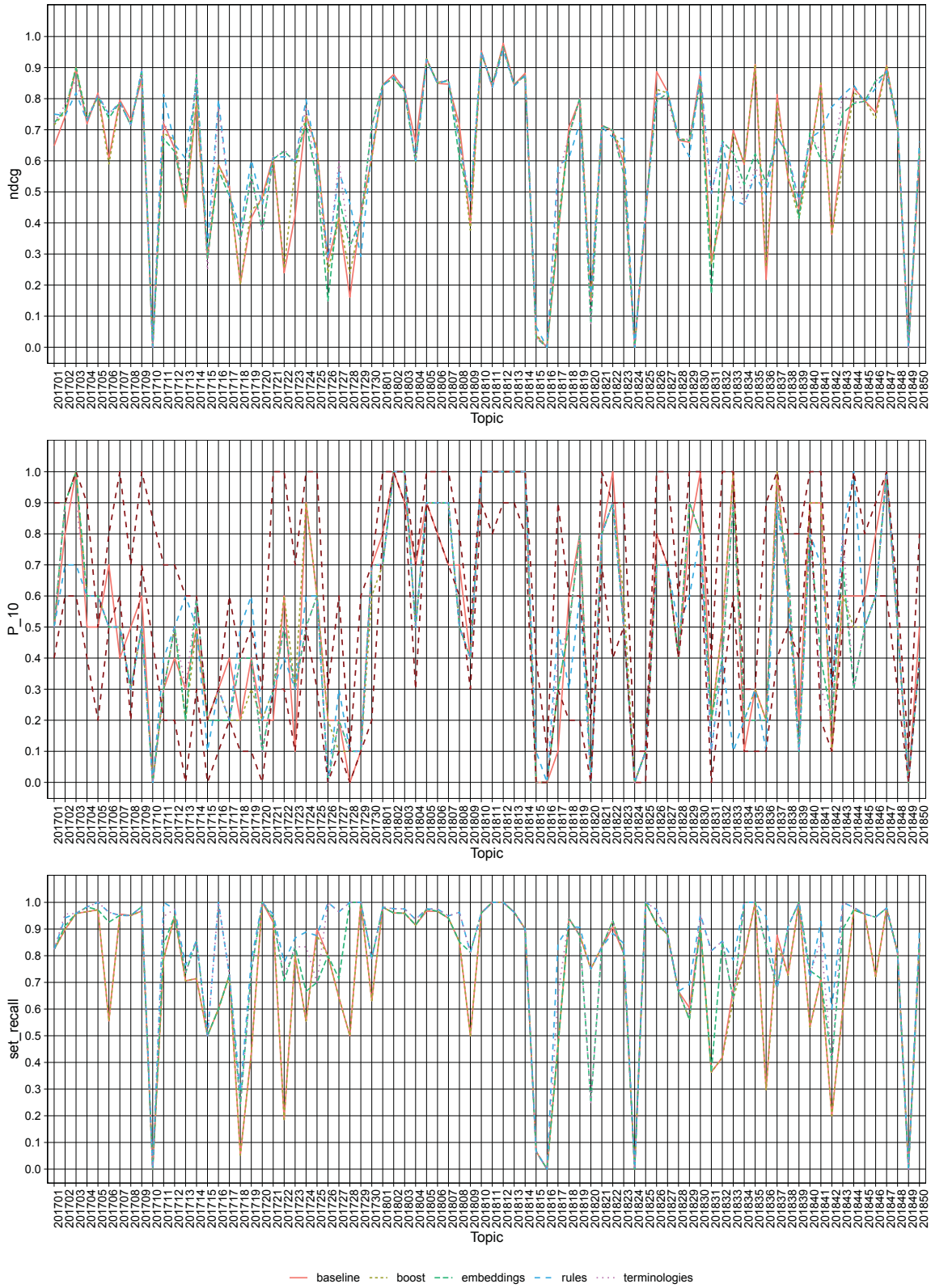

Figure A.2: Clinical trials: NDCG, P@10, and recall per topic.

BIBLIOGRAPHY

1. Oleynik M, Kugic A, Kasác Z, and Kreuzthaler M. Evaluating Shallow and Deep Learning Strategies for the 2018 n2c2 Shared Task on Clinical Text Classification. JAMIA 2019;26:1247–54 (cit. on pp. ix, 40, 41, 50).
2. U.S. National Research Council. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC, USA: The National Academies Press, 2011 (cit. on p. 1).
3. Collins FS and Varmus H. A New Initiative on Precision Medicine. N. Engl. J. Med. 2015;372:793–5 (cit. on p. 1).
4. Frey LJ, Bernstam EV, and Denny JC. Precision Medicine Informatics. JAMIA 2016;23:668–70 (cit. on p. 1).
5. Landoni G, Comis M, Conte M, Finco G, Mucchetti M, Paternoster G, et al. Mortality in Multicenter Critical Care Trials: an Analysis of Interventions with a Significant Effect. Critical care medicine 2015;43:1559–68 (cit. on p. 1).
6. Ospina-Tascón GA, Büchele GL, and Vincent JL. Multicenter, Randomized, Controlled Trials Evaluating Mortality in Intensive Care: Doomed to Fail? Critical care medicine 2008;36:1311–22 (cit. on p. 1).
7. Safran C, Bloomrosen M, Hammond WE, Labkoff SE, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. JAMIA 2007;14:1–9 (cit. on p. 1).
8. Geneletti S, Richardson S, and Best N. Adjusting for Selection Bias in Retrospective, Case–Control Studies. Biostatistics. 10:17–31 (cit. on pp. 1, 35).
9. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, Shirtcliffe P, et al. External Validity of Randomised Controlled Trials in Asthma: to Whom do the Results of the Trials Apply? Thorax 2007;62:219–23 (cit. on p. 1).
10. Naumann T. Leveraging text representations for clinical predictive tasks. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA, 2018 (cit. on p. 1).
11. Horsky J, Drucker EA, and Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. In: *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4-8, 2017*. AMIA, 2017 (cit. on pp. 2, 35).

12. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep Learning in Clinical Natural Language Processing: a Methodical Review. JAMIA 2020;27:457–70 (cit. on pp. 2, 5, 8).
13. Manning CD and Schütze H. Foundations of Statistical Natural Language Processing. MIT Press, 1999 (cit. on pp. 2, 7, 9, 15, 16, 61).
14. Meystre SM, Savova GK, Kipper-Schuler KC, and Hurdle JF. Extracting Information from Textual Documents in the Electronic Health Record: a Review of Recent Research. Yearb. Med. Inform. 2008;17:128–44 (cit. on p. 2).
15. Goldberg Y. Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2017 (cit. on p. 2).
16. Weber GM, Mandl KD, and Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA 2014;311:2479–80 (cit. on p. 2).
17. Alpaydin E. Introduction to Machine Learning. Adaptive computation and machine learning. MIT Press, 2004 (cit. on p. 5).
18. Chauvin Y and Rumelhart DE, eds. Backpropagation: Theory, Architectures, and Applications. Psychology Press, 1995 (cit. on p. 5).
19. Hahn U and Oleynik M. Medical Information Extraction in the Age of Deep Learning (to appear). Yearb. Med. Inform. 2020 (cit. on p. 5).
20. Xiao C, Choi E, and Sun J. Opportunities and Challenges in Developing Deep Learning Models Using Electronic Health Records Data: a Systematic Review. JAMIA 2018;25:1419–28 (cit. on p. 5).
21. Shickel B, Tighe P, Bihorac A, and Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J. Biomed. Health Informatics 2018;22:1589–604 (cit. on p. 5).
22. Pan SJ and Yang Q. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 2010;22:1345–59 (cit. on p. 6).
23. Raina R, Battle A, Lee H, Packer B, and Ng AY. Self-Taught Learning: Transfer Learning from Unlabeled Data. In: *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, OR, USA, June 20-24, 2007*. Ed. by Ghahramani Z. Vol. 227. ACM International Conference Proceeding Series. ACM, 2007:759–66 (cit. on p. 6).
24. Bengio Y, Courville AC, and Vincent P. Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013;35:1798–828 (cit. on p. 6).

25. Goodfellow IJ, Bengio Y, and Courville AC. Deep Learning. Adaptive computation and machine learning. MIT Press, 2016 (cit. on pp. 6, 43).

26. Kalyan KS and Sangeetha S. SECNLP: A Survey of Embeddings in Clinical Natural Language Processing. J. Biomed. Informatics 2020;101:103323 (cit. on pp. 6, 8).

27. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, and Rudzicz F. A Survey of Word Embeddings for Clinical Text. J. Biomed. Informatics X 2019;4:100057 (cit. on pp. 6, 8).

28. Pakhomov SVS, McInnes BT, Adam T, Liu Y, Pedersen T, and Melton GB. Semantic Similarity and Relatedness between Clinical Terms: an Experimental Study. In: *AMIA 2010, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 13-17, 2010*. Ed. by Kuperman GJ, Friedman C, and Sittig DF. AMIA, 2010:572–6 (cit. on p. 6).

29. Pakhomov SVS, Pedersen T, McInnes BT, Melton GB, Ruggieri A, and Chute CG. Towards a Framework for Developing Semantic Relatedness Reference Standards. J. Biomed. Informatics 2011;44:251–65 (cit. on p. 6).

30. Peng Y, Yan S, and Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*. Ed. by Demner-Fushman D, Cohen KB, Ananiadou S, and Tsujii J. Association for Computational Linguistics, 2019:58–65 (cit. on pp. 6, 75).

31. Chiu B, Crichton GKO, Korhonen A, and Pyysalo S. How to Train Good Word Embeddings for Biomedical NLP. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2016, Berlin, Germany, August 12, 2016*. Ed. by Cohen KB, Demner-Fushman D, Ananiadou S, and Tsujii J. Association for Computational Linguistics, 2016:166–74 (cit. on p. 6).

32. Levy O, Goldberg Y, and Dagan I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. Trans. Assoc. Comput. Linguistics 2015;3:211–25 (cit. on p. 6).

33. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, and Salakhutdinov R. Dropout: a Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 2014;15:1929–58 (cit. on p. 6).

34. Riley MD. Some Applications of Tree-Based Modelling to Speech and Language. In: *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, MA, USA, HLT 1989, October 15-18, 1989*. ACL, 1989 (cit. on p. 7).

35. Kreuzthaler M and Schulz S. Detection of Sentence Boundaries and Abbreviations in Clinical Narratives. BMC Med. Inf. & Decision Making 2015;15:S4 (cit. on p. 7).

36. Kučera H and Francis WN. Computational Analysis of Present-Day American English. Dartmouth Publishing Group, 1967 (cit. on p. 7).

37. Salton G. The SMART Retrieval System—Experiments in Automatic Document Processing. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971 (cit. on p. 8).

38. Mikolov T, Chen K, Corrado G, and Dean J. Efficient Estimation of Word Representations in Vector Space. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Bengio Y and LeCun Y. 2013 (cit. on p. 8).

39. Mikolov T, Sutskever I, Chen K, Corrado GS, and Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, NE, USA*. Ed. by Burges CJC, Bottou L, Ghahramani Z, and Weinberger KQ. 2013:3111–9 (cit. on pp. 8, 28).

40. Pennington J, Socher R, and Manning CD. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Moschitti A, Pang B, and Daelemans W. ACL, 2014:1532–43 (cit. on p. 8).

41. Bojanowski P, Grave E, Joulin A, and Mikolov T. Enriching Word Vectors with Subword Information. TACL 2017;5:135–46 (cit. on p. 8).

42. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, and Vollgraf R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*. Ed. by Ammar W, Louis A, and Mostafazadeh N. Association for Computational Linguistics, 2019:54–9 (cit. on p. 8).

43. Pyysalo S, Ginter F, Moen H, Salakoski T, and Ananiadoua S. Distributional Semantics Resources for Biomedical Text Processing. In: *Proceedings of 5th International Symposium on Languages in Biology and Medicine, LBM 2013, Tokyo, Japan, December 12-13, 2013*. 2013:39–44 (cit. on pp. 8, 14).

44. Zhang Y, Chen Q, Yang Z, Lin H, and Lu Z. BioWordVec, Improving Biomedical Word Embeddings with Subword Information and MeSH. Scientific data 2019;6:52 (cit. on p. 9).

45. Chen Q, Peng Y, and Lu Z. BioSentVec: Creating Sentence Embeddings for Biomedical Texts. In: *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019, Xi'an, China, June 10-13, 2019*. IEEE, 2019:1–5 (cit. on pp. 9, 14, 37, 38, 40).

46. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. J. Biomed. Informatics 2018;87:12–20 (cit. on p. 9).

47. Si Y, Wang J, Xu H, and Roberts KE. Enhancing Clinical Concept Extraction with Contextual Embeddings. JAMIA 2019;26:1297–304 (cit. on p. 9).

48. Roberts K. Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP. In: *Proceedings of the Clinical Natural Language Processing Workshop, ClinicalNLP@COLING 2016, Osaka, Japan, December 11, 2016*. Ed. by Rumshisky A, Roberts K, Bethard S, and Naumann T. The COLING 2016 Organizing Committee, 2016:54–63 (cit. on pp. 9, 37).

49. Navigli R. Word Sense Disambiguation: A Survey. ACM Comput. Surv. 2009;41:10:1–10:69 (cit. on p. 9).

50. Yarowsky D. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992*. 1992:454–60 (cit. on p. 9).

51. Yarowsky D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: *33rd Annual Meeting of the Association for Computational Linguistics, 26-30 June 1995, MIT, Cambridge, MA, USA, Proceedings*. Ed. by Uszkoreit H. Morgan Kaufmann Publishers / ACL, 1995:189–96 (cit. on p. 9).

52. Gale WA, Church KW, and Yarowsky D. A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities 1992;26:415–39 (cit. on p. 10).

53. Brown PF, Pietra SD, Pietra VJD, and Mercer RL. Word-Sense Disambiguation Using Statistical Methods. In: *29th Annual Meeting of the Association for Computational Linguistics, 18-21 June 1991, University of California, Berkeley, CA, USA, Proceedings*. Ed. by Appelt DE. ACL, 1991:264–70 (cit. on p. 10).

54. Choueka Y and Lusignan S. Disambiguation by Short Contexts. Computers and the Humanities 1985;19:147–57 (cit. on p. 10).

55. Gale WA, Church KW, and Yarowsky D. Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. In: *30th Annual Meeting of the Association for Computational Linguistics, 28 June - 2 July 1992, University of Delaware, Newark, DE, USA, Proceedings*. Ed. by Thompson HS. ACL, 1992:249–57 (cit. on p. 10).

56. Sanderson M and Rijsbergen CJ van. The Impact on Retrieval Effectiveness of Skewed Frequency Distributions. ACM Trans. Inf. Syst. 1999;17:440–65 (cit. on p. 10).

57. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova GK, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (Valencia, Spain). Ed. by Forner P, Müller H, Paredes R, Rosso P, and Stein B. Vol. 8138. Lecture Notes in Computer Science. Springer, 2013:212–31 (cit. on pp. 10, 22).

58. Mowery DL, South BR, Christensen LM, Leng J, Peltonen L, Salanterä S, et al. Normalizing Acronyms and Abbreviations to Aid Patient Understanding of Clinical Texts: ShARe/CLEF eHealth Challenge 2013, Task 2. J. Biomedical Semantics 2016;7:43 (cit. on pp. 10, 21, 22).

59. Manning CD, Raghavan P, and Schütze H. An Introduction to Information Retrieval. Cambridge University Press, 2008 (cit. on pp. 10–12, 16, 17, 54).

60. Maron ME and Kuhns JL. On Relevance, Probabilistic Indexing and Information Retrieval. J. ACM 1960;7:216–44 (cit. on p. 10).

61. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*. Ed. by Nedellec C and Rouveirol C. Vol. 1398. Lecture Notes in Computer Science. Springer, 1998:137–42 (cit. on pp. 10, 42).

62. Berkson J. Application of the Logistic Function to Bio-Assay. J. Am. Stat. Assoc. 1944;39:357–65 (cit. on p. 10).

63. Joulin A, Grave E, Bojanowski P, and Mikolov T. Bag of Tricks for Efficient Text Classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. Ed. by Lapata M, Blunsom P, and Koller A. Association for Computational Linguistics, 2017:427–31 (cit. on pp. 10, 38).

64. Chiticariu L, Li Y, and Reiss FR. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, WA, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2013:827–32 (cit. on p. 11).

65. Liu T. Learning to Rank for Information Retrieval. Found. Trends Inf. Retr. 2009;3:225–331 (cit. on p. 12).

66. Hersh WR. Information Retrieval and Digital Libraries. In: *Biomedical Informatics*. Ed. by Shortliffe EH and Cimino JJ. Springer, 2014. Chap. 19:613–41 (cit. on pp. 13, 14).

67. He J and Li K. How Comprehensive is the PubMed Central Open Access Full-Text Database? In: *iConference 2019 Proceedings*. iSchools, 2019 (cit. on p. 13).

68. Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data 2016;3:160035 (cit. on p. 13).

69. Hand DJ. Assessing the Performance of Classification Methods. International Statistical Review 2012;80:400–14 (cit. on p. 14).

70. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural computation 1998;10:1895–923 (cit. on p. 17).

71. Smucker MD, Allan J, and Carterette B. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*. Ed. by Silva MJ, Laender AHF, Baeza-Yates RA, McGuinness DL, Olstad B, Olsen ØH, et al. ACM, 2007:623–32 (cit. on p. 17).

72. Xu H, Stetson PD, and Friedman C. A Study of Abbreviations in Clinical Notes. In: *AMIA 2007, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 10-14, 2007*. AMIA, 2007 (cit. on p. 19).

73. Schwartz AS and Hearst MA. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. In: *Proceedings of the 8th Pacific Symposium on Biocomputing, PSB 2003, Lihue, HI, USA, January 3-7, 2003*. Ed. by Altman RB, Dunker AK, Hunter L, and Klein TE. 2003:451–62 (cit. on p. 19).

74. Sheppard JE, Weidner LCE, Zakai S, Fountain-Polley S, and Williams J. Ambiguous Abbreviations: an Audit of Abbreviations in Paediatric Note Keeping. Arch. Dis. Child. 2008;93:204–6 (cit. on p. 19).

75. Walsh KE and Gurwitz JH. Medical Abbreviations: Writing Little and Communicating Less. Arch. Dis. Child. 2008;93:816–7 (cit. on p. 19).

76. Grossman LV, Creber RMM, Restaino S, and Vawdrey DK. Sharing Clinical Notes with Hospitalized Patients via an Acute Care Portal. In: *AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4-8, 2017*. AMIA, 2017 (cit. on p. 19).

77. Spasic I. Acronyms as an Integral Part of Multi-Word Term Recognition - A Token of Appreciation. IEEE Access 2018;6:8351–63 (cit. on pp. 20, 21, 23).

78. Yu H, Hripcsak G, and Friedman C. Mapping Abbreviations to Full Forms in Biomedical Articles. JAMIA 2002;9:262–72 (cit. on p. 20).

79. Zhou W, Torvik VI, and Smalheiser NR. ADAM: Another Database of Abbreviations in MEDLINE. Bioinform. 2006;22:2813–8 (cit. on p. 20).

80. Ao H and Takagi T. ALICE: An Algorithm to Extract Abbreviations from MEDLINE. JAMIA 2005;12:576–86 (cit. on p. 20).

81. Firth JR. A Synopsis of Linguistic Theory, 1930-1955. In: *Studies in Linguistic Analysis*. Vol. 1952-59. Oxford: Basil Blackwell, 1962:1–32 (cit. on p. 21).

82. Baroni M, Dinu G, and Kruszewski G. Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. The Association for Computer Linguistics, 2014:238–47 (cit. on p. 21).

83. Liu H, Lussier YA, and Friedman C. Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: An Unsupervised Method. J. Biomed. Informatics 2001;34:249–61 (cit. on p. 21).

84. Pakhomov SVS. Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002:160–7 (cit. on pp. 21, 22).

85. Wu Y, Tang B, Jiang M, Moon S, Denny JC, and Xu H. Clinical Acronym/Abbreviation Normalization using a Hybrid Approach. In: *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*. Ed. by Forner P, Navigli R, Tufis D, and Ferro N. Vol. 1179. CEUR Workshop Proceedings. CEUR-WS.org, 2013 (cit. on pp. 21, 22).

86. Siklósi B, Novák A, and Prószéky G. Resolving Abbreviations in Clinical Texts Without Pre-Existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, BioTxtM@LREC 2014, Harpa, Iceland, May 31, 2014*. Ed. by Ananiadou S, Choukri K, Cohen KB, Demner-Fushman D, Hajic J, Hanbury A, et al. European Language Resources Association (ELRA), 2014:69–75 (cit. on pp. 21, 22).

87. Kirchhoff K and Turner AM. Unsupervised Resolution of Acronyms and Abbreviations in Nursing Notes Using Document-Level Context Models. In: *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2016, Austin, TX, USA, November 5, 2016*. Ed. by Grouin C, Hamon T, Névéol A, and Zweigenbaum P. Association for Computational Linguistics, 2016:52–60 (cit. on pp. 21, 22).

88. Charbonnier J and Wartena C. Using Word Embeddings for Unsupervised Acronym Disambiguation. In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, NM, USA, August 20-26, 2018*. Ed. by Bender EM, Derczynski L, and Isabelle P. Association for Computational Linguistics, 2018:2610–9 (cit. on pp. 21, 22).

89. León FS. ARBOREx: Abbreviation Resolution Based on Regular Expressions for BARR2. In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*. Ed. by Rosso P, Gonzalo J, Martínez R, Montalvo S, and Albornoz JC de. Vol. 2150. CEUR Workshop Proceedings. CEUR-WS.org, 2018:302–15 (cit. on pp. 21, 23).

90. Intxaurrondo A, Marimon M, Gonzalez-Agirre A, López-Martín JA, Rodriguez H, Santamaría J, et al. Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results. In: *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*. Ed. by Rosso P, Gonzalo J, Martínez R, Montalvo S, and Albornoz JC de. Vol. 2150. CEUR Workshop Proceedings. CEUR-WS.org, 2018:280–9 (cit. on p. 23).

91. Park Y and Byrd RJ. Hybrid Text Mining for Finding Abbreviations and their Definitions. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2001, Pittsburgh, PA, USA, June 3-4, 2001*. ACL, 2001 (cit. on p. 24).

92. Cohn T. Performance Metrics for Word Sense Disambiguation. In: *Proceedings of the Australasian Language Technology Workshop, ALTA 2003, Melbourne, Australia, December 8-12, 2003*. Ed. by Bow C and Hughes B. Australasian Language Technology Association, 2003:86–93 (cit. on p. 25).

93. Pomares-Quimbaya A, López-Úbeda P, Oleynik M, and Schulz S. Leveraging PubMed to create a Specialty-Based Sense Inventory for Spanish Acronym Resolution (to appear). In: *MIE 2020: Proceedings of the 30th Medical Informatics Europe*. Studies in Health Technology and Informatics. IOS Press, 2020 (cit. on p. 33).

94. Mann CJ. Observational Research Methods. Research Design II: Cohort, Cross Sectional, and Case-Control Studies. Emergency Medicine Journal. 20:54–60 (cit. on p. 35).

95. Hebal F, Nanney E, Stake C, Miller M, Lales G, and Barsness KA. Automated Data Extraction: Merging Clinical Care with Real-Time Cohort-Specific Research and Quality Improvement Data. Journal of Pediatric Surgery. 52:149–52 (cit. on p. 36).

96. Shivade CP, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A Review of Approaches to Identifying Patient Phenotype Cohorts Using Electronic Health Records. JAMIA 2014;21:221–30 (cit. on p. 36).

97. Hripcsak G and Albers DJ. Next-Generation Phenotyping of Electronic Health Records. JAMIA 2013;20:117–21 (cit. on p. 36).

98. Pathak J, Kho AN, and Denny JC. Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives. JAMIA e2. 20:e206–e211 (cit. on p. 36).

99. Stanfill MH, Williams M, Fenton SH, Jenders RA, and Hersh WR. A Systematic Literature Review of Automated Clinical Coding and Classification Systems. JAMIA 2010;17:646–51 (cit. on p. 36).

100. Shaikh GM, Shuib NLM, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues. Expert Syst. Appl. 2019;116:494–520 (cit. on p. 36).

101. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, and Calster BV. A Systematic Review shows no Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. J. Clin. Epidemiol. 2019;110:12–22 (cit. on p. 36).

102. Wilcox AB and Hripcsak G. Classification Algorithms Applied to Narrative Reports. In: *AMIA 1999, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 6-10, 1999*. AMIA, 1999 (cit. on p. 36).

103. Jouhet V, Defossez G, Burgun A, Le PB, Levillain P, Ingrand P, et al. Automated Classification of Free-Text Pathology Reports for Registration of Incident Cases of Cancer. Methods Inf Med. 51:242–51 (cit. on p. 36).

104. Lipton ZC, Kale DC, Elkan C, and Wetzel RC. Learning to Diagnose with LSTM Recurrent Neural Networks. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Bengio Y and LeCun Y. 2016 (cit. on pp. 36, 43).

105. Yao L, Mao C, and Luo Y. Clinical Text Classification with Rule-Based Features and Knowledge-Guided Convolutional Neural Networks. BMC Med. Inf. & Decision Making 2019;19-S:31–9 (cit. on p. 36).

106. Karimi S, Dai X, Hassanzadeh H, and Nguyen AN. Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods. In: *BioNLP 2017, Vancouver, Canada, August 4, 2017*. Ed. by Cohen KB, Demner-Fushman D, Ananiadou S, and Tsujii J. Association for Computational Linguistics, 2017:328–32 (cit. on p. 37).

107. Stubbs A, Kotfila C, Xu H, and Uzuner Ö. Identifying Risk Factors for Heart Disease Over Time: Overview of 2014 i2b2/UTHealth Shared Task Track 2. J. Biomed. Informatics 2015;58:S67–S77 (cit. on pp. 37, 51).

108. Kotfila C and Uzuner Ö. A Systematic Comparison of Feature Space Effects on Disease Classifier Performance for Phenotype Identification of Five Diseases. J. Biomed. Informatics 2015;58:S92–S102 (cit. on p. 37).

109. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, and Demner-Fushman D. The Role of Fine-Grained Annotations in Supervised Recognition of Risk Factors for Heart Disease from EHRs. J. Biomed. Informatics 2015;58:S111–S119 (cit. on p. 37).

110. Hochreiter S and Schmidhuber J. Long Short-Term Memory. Neural Computation 1997;9:1735–80 (cit. on p. 43).

111. Gao S, Young MT, Qiu JX, Yoon H, Christian JB, Fearn PA, et al. Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports. JAMIA 2018;25:321–30 (cit. on p. 43).

112. Jagannatha AN and Yu H. Bidirectional RNN for Medical Event Detection in Electronic Health Records. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, June 12-17, 2016*. Ed. by Knight K, Nenkova A, and Rambow O. The Association for Computational Linguistics, 2016:473–82 (cit. on p. 43).

113. Kingma DP and Ba J. Adam: A Method for Stochastic Optimization. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio Y and LeCun Y. 2015 (cit. on p. 43).

114. Stubbs A, Filannino M, Soysal E, Henry S, and Uzuner Ö. Cohort Selection for Clinical Trials: n2c2 2018 Shared Task Track 1. JAMIA 2019;26:1163–71 (cit. on p. 49).

115. Vydiswaran VGV, Strayhorn A, Zhao X, Robinson P, Agarwal M, Bagazinski E, et al. Hybrid Bag of Approaches to Characterize Selection Criteria for Cohort Identification. JAMIA 2019;26:1172–80 (cit. on pp. 49–51).

116. Tannier X, Paris N, Cisneros H, Daniel C, Doutreligne M, Duclos C, et al. Hybrid Approaches for our Participation to the n2c2 Challenge on Cohort Selection for Clinical Trials. 2019. arXiv: 1903.07879 (cit. on pp. 49–51).

117. Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, et al. Clinical Trial Cohort Selection Based on Multi-Level Rule-Based Natural Language Processing System. JAMIA 2019;26:1218–26 (cit. on pp. 49, 50).

118. Ni Y, Bermudez M, Kennebeck S, Liddy-Hicks S, and Dexheimer J. A Real-Time Automated Patient Screening System for Clinical Trials Eligibility in an Emergency Department: Design and Evaluation. JMIR 2019;7:e14185 (cit. on p. 50).

119. Rawal S, Prakash A, Adhya S, Kulkarni S, Anwar S, Baral C, et al. Developing and Using Special-Purpose Lexicons for Cohort Selection from Clinical Notes. 2019. arXiv: 1902.09674 (cit. on p. 50).

120. Karystianis G, Florez-Vargas O, Butler T, and Nenadic G. A Rule-Based Approach to Identify Patient Eligibility Criteria for Clinical Trials from Narrative Longitudinal Records. JAMIA Open 2019;2:521–7 (cit. on p. 50).

121. Xiong Y, Shi X, Chen S, Jiang D, Tang B, Wang X, et al. Cohort Selection for Clinical Trials using Hierarchical Neural Network. JAMIA 2019;26:1203–8 (cit. on pp. 50, 51).

122. Shi J, Graves K, and Hurdle JF. A Generic Rule-Based System for Clinical Trial Patient Selection. 2019. arXiv: 1907.06860 (cit. on p. 50).

123. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, and Brown SH. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. JAMIA 2006;13:277–88 (cit. on p. 54).

124. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, and Brown SH. Model Formulation: A Model for Evaluating Interface Terminologies. JAMIA 2008;15:65–76 (cit. on p. 54).

125. Roy D, Paul D, Mitra M, and Garain U. Using Word Embeddings for Automatic Query Expansion. 2016. arXiv: 1606.07608 (cit. on p. 54).

126. Kuzi S, Shtok A, and Kurland O. Query Expansion Using Word Embeddings. In: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*. Ed. by Mukhopadhyay S, Zhai C, Bertino E, Crestani F, Mostafa J, Tang J, et al. ACM, 2016:1929–32 (cit. on p. 54).

127. Voorhees EM and Tong RM. Overview of the TREC 2011 Medical Records Track. In: *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, MD, USA, November 15-18, 2011*. Ed. by Voorhees EM and Buckland LP. Vol. 500-296. NIST Special Publication. National Institute of Standards and Technology (NIST), 2011 (cit. on p. 54).

128. Voorhees EM and Hersh WR. Overview of the TREC 2012 Medical Records Track. In: *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, MD, USA, November 6-9, 2012*. Ed. by Voorhees EM and Buckland LP. Vol. 500-298. NIST Special Publication. National Institute of Standards and Technology (NIST), 2012 (cit. on p. 54).

129. Edinger T, Cohen AM, Bedrick S, Ambert KH, and Hersh WR. Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track. In: *AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 3-7, 2012*. AMIA, 2012 (cit. on p. 54).

130. Simpson MS, Voorhees EM, and Hersh WR. Overview of the TREC 2014 Clinical Decision Support Track. In: *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, MD, USA, November 19-21, 2014*. Ed. by Voorhees EM and Ellis A. Vol. 500-308. NIST Special Publication. National Institute of Standards and Technology (NIST), 2014 (cit. on p. 55).

131. Roberts K, Simpson MS, Voorhees EM, and Hersh WR. Overview of the TREC 2015 Clinical Decision Support Track. In: *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, MD, USA, November 17-20, 2015*. Ed. by Voorhees EM and Ellis A. Vol. 500-319. NIST Special Publication. National Institute of Standards and Technology (NIST), 2015 (cit. on p. 55).

132. Roberts K, Demner-Fushman D, Voorhees EM, and Hersh WR. Overview of the TREC 2016 Clinical Decision Support Track. In: *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, MD, USA, November 15-18, 2016*. Ed. by Voorhees EM and Ellis A. Vol. 500-321. NIST Special Publication. National Institute of Standards and Technology (NIST), 2016 (cit. on p. 55).

133. Goodwin T and Harabagiu SM. UTD at TREC 2014: Query Expansion for Clinical Decision Support. Tech. rep. 2014 (cit. on p. 55).

134. Nguyen V, Karimi S, Falamaki S, and Paris C. Benchmarking Clinical Decision Support Search. 2018. arXiv: 1801.09322 (cit. on p. 55).

135. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2017 Precision Medicine Track. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, MD, USA, November 15-17, 2017*. Ed. by Voorhees EM and Ellis A. Vol. 500-324. NIST Special Publication. National Institute of Standards and Technology (NIST), 2017 (cit. on pp. 55, 57, 70).

136. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, and Lazar AJ. Overview of the TREC 2018 Precision Medicine Track. In: *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, MD, USA, November 14-16, 2018*. Ed. by Voorhees EM and Ellis A. Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology (NIST), 2018 (cit. on pp. 55, 58, 70).

137. Nguyen V, Karimi S, Falamaki S, Aliod DM, Paris C, and Wan S. CSIRO at 2017 TREC Precision Medicine Track. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, MD, USA, November 15-17, 2017*. Ed. by Voorhees EM and Ellis A. Vol. 500-324. NIST Special Publication. National Institute of Standards and Technology (NIST), 2017 (cit. on p. 55).

138. Eghlidi NF, Griner J, Mesot N, Werra L von, and Eickhoff C. ETH Zurich at TREC Precision Medicine 2017. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, MD, USA, November 15-17, 2017*. Ed. by Voorhees EM and Ellis A. Vol. 500-324. NIST Special Publication. National Institute of Standards and Technology (NIST), 2017 (cit. on p. 55).

139. Nishani L, Kolla M, and Baruah G. KlickLabs at TREC 2018 Precision Medicine Track. In: *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, MD, USA, November 14-16, 2018*. Ed. by Voorhees EM and Ellis A. Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology (NIST), 2018 (cit. on p. 55).

140. Baruah P, Dulepet R, Qian K, and Eickhoff C. Brown University at TREC Precision Medicine 2018. In: *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, MD, USA, November 14-16, 2018*. Ed. by Voorhees EM and Ellis A. Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology (NIST), 2018 (cit. on p. 55).

141. Diaz F, Mitra B, and Craswell N. Query Expansion with Locally-Trained Word Embeddings. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016 (cit. on p. 56).

142. Rattinger A, Goff JL, and Guetl C. Local Word Embeddings for Query Expansion Based on Co-Authorship and Citations. In: *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2018) co-located with the 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018*. Ed. by Mayr P, Frommholz I, and Cabanac G. Vol. 2080. CEUR Workshop Proceedings. CEUR-WS.org, 2018:46–53 (cit. on p. 56).

143. Yilmaz E, Kanoulas E, and Aslam JA. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*. Ed. by Myaeng S, Oard DW, Sebastiani F, Chua T, and Leong M. ACM, 2008:603–10 (cit. on p. 59).

144. Saito T and Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 2015;10:1–21 (cit. on p. 59).

145. Jones KS, Walker S, and Robertson SE. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments - Part 2. Inf. Process. Manag. 2000;36:809–40 (cit. on p. 61).

146. Ghawi R and Pfeffer J. Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity. Open Comput. Sci. 2019;9:160–80 (cit. on p. 68).

147. Faessler E, Oleynik M, and Hahn U. What makes a Top-Performing Precision Medicine Search Engine? Tracing Main System Features in a Systematic Way (to appear). In: *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Xi'an, China, July 25-30, 2020*. ACM, 2020 (cit. on p. 68).

148. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2019 Precision Medicine Track. In: *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, MD, USA, November 13-15, 2019*. Ed. by Voorhees EM and Ellis A. Vol. 1250. NIST Special Publication. National Institute of Standards and Technology (NIST), 2019 (cit. on p. 70).

149. Devlin J, Chang M, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Burstein J, Doran C, and Solorio T. Association for Computational Linguistics, 2019:4171–86 (cit. on p. 74).

150. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. arXiv: 1907.11692 (cit. on p. 74).

151. Clark K, Luong MT, Le QV, and Manning CD. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2020 (cit. on p. 74).

152. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Ed. by Guyon I, Luxburg U von, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, et al. 2017:5998–6008 (cit. on p. 74).

153. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinform. 2020;36:1234–40 (cit. on p. 75).

154. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, MN, USA: Association for Computational Linguistics, 2019:72–8 (cit. on p. 75).

155. Huang K, Altosaar J, and Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2019. arXiv: 1904.05342 (cit. on p. 75).

156. Martin L, Müller B, Suárez PJO, Dupont Y, Romary L, Clergerie ÉV de la, et al. CamemBERT: a Tasty French Language Model. 2019. arXiv: 1911.03894 (cit. on p. 75).

157. Ribeiro MT, Singh S, and Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, and Rastogi R. ACM, 2016:1135–44 (cit. on p. 75).

158. Camacho-Collados J and Pilehvar MT. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. J. Artif. Intell. Res. 2018;63:743–88 (cit. on p. 75).

159. Zhang Z, Han X, Liu Z, Jiang X, Sun M, and Liu Q. ERNIE: Enhanced Language Representation with Informative Entities. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Korhonen A, Traum DR, and Màrquez L. Association for Computational Linguistics, 2019:1441–51 (cit. on p. 75).

INDEX

abbreviation, 2, 20
accuracy, 15
acronym, 2, 20
    detection, 20
    disambiguation, 20
    expansion, 20

backpropagation, 5
baseline, 16, 27, 41, 61
BioWordVec, 9, 14, 40

cohort selection, 35
continuous bag of words, 6, 8, 28
convolutional neural network, 10
count-based model, 21

deep learning, 5, 41
dis_max, 60
distributional hypothesis, 12, 21
DL, *see* deep learning
domain adaptation, 6
dropout, 6

evaluation
    extrinsic, 6
    intrinsic, 6

f-score, 16, 40
fastText, 8, 33, 40
Fisher's randomization test, 17, 59

hypernymy, 12, 53, 60

i2b2, 23
inferred normalized discounted cumulative gain, 59
infNDCG, *see* inferred normalized discounted cumulative gain
information retrieval, 11, 54
inverted index, 11
IR, *see* information retrieval

kappa index, 16, 29

logistic regression, 42
long short-term memory, 11, 43
lowercasing, 7, 41, 60
LR, *see* logistic regression
LSTM, *see* long short-term memory

machine learning
    supervised, 5, 9
    unsupervised, 5, 10
McNemar's test, 17, 26, 41
MEDLINE, 13
meronymy, 12, 53
MIMIC-III, 13, 22, 34, 40
ML, *see* machine learning
multi-task learning, 6

n-gram, 27
natural language processing, 2, 7
NDCG, *see* normalized discounted cumulative gain
neural network, 5
NLP, *see* natural language processing
NN, *see* neural network
normalized discounted cumulative gain, 59

one-hot encoding, 8

patient phenotyping, 36
perceptron network, 43
PMC, 13
precision, 15, 25, 40, 59
precision medicine, 1, 56

precision-recall curve, 59
precision-recall tradeoff, 12
predictive model, 21
PubMed, 1, 13, 20, 23, 40, 55, 56, 62

query boosting, 61
query expansion, 12, 62

recall, 15, 25, 40, 59
recurrent neural network, 11
RNN, *see* recurrent neural network
rules, 11, 26, 41, 62

self-taught learning, 6
sense discrimination, 10
sense inventory, 20, 27
sentence detection, 7, 41
shallow learning, 10, 41
skip-gram, 6, 8
stopword, 8, 41, 60
support vector machine, 22, 36, 42
SVM, *see* support vector machine
synonymy, 12, 53, 60

term conflation, 20
terminology, 12, 64
  interface, 54
  reference, 54
text classification, 10, 41
text cleansing, 7, 26, 41
tf-idf, 11, 36, 42, 61
tokenization, 7, 41
transfer learning, 2, 5

UMLS, 20, 21
unsupervised pre-training, 6

word embeddings, 6, 8, 27, 54, 62
word2vec, 6, 8, 21, 28, 62

COLOPHON

This document was typeset in LaTeX using the typographical look-and-feel `classicthesis`. Most of the plots in this thesis were generated using `R` and `ggplot2`. The bibliography was typeset using `biber`.

Final Version as of June 8, 2020 (version RC3).