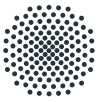


Universität Stuttgart

Deep Learning aided Clinical Decision Support

Von der Fakultät 5 - Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

Rudolf Schneider

aus Ribnitz-Damgarten

Hauptberichter:	Prof. Dr. Steffen Staab
Mitberichter:	Prof. Dr. Alexander Löser
Mitberichter:	Prof. Dr. Georg Rehm

Tag der mündlichen Prüfung: 30.06.2023

Institut für Parallele und Verteilte Systeme

2023

Publications

The following publications form the basis of this thesis[1]:

1. **R. Schneider**, C. Guder, T. Kilias, A. Löser, J. Graupmann, and O. Kozachuk, "Interactive Relation Extraction in Main Memory Database Systems," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, 2016, vol. 26, pp. 103–106.
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework with support from T. Kilias.
   - R. Schneider led the verification of the analytical methods with input from T. Kilias.
   - R. Schneider led the implementation of the theory, baselines, and execution of computations and was assisted by C. Guder.
   - R. Schneider was steering various in-depth discussions with T. Kilias, A. Löser, J. Graupmann, and O. Kozachuk that led to the final experiments and results.
   - All authors discussed the results and contributed to the final manuscript.

[1] In order of date of publication

2. **R. Schneider**, T. Oberhauser, T. Klatt, F. A. Gers, and A. Löser, "Analysing errors of open information extraction systems," presented at the Building Linguistically Generalizable NLP Systems Workshop at EMNLP 2017, Copenhagen, Denmark, 2017.
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework.
   - R. Schneider led the verification of the analytical methods with input from F. A. Gers, and A. Löser.
   - R. Schneider led the implementation of the theory, baselines, and execution of computations and was assisted by T. Oberhauser and T. Klatt.
   - R. Schneider was steering various in-depth discussions with F. A. Gers, and A. Löser that led to the final experiments and results.
   - All authors discussed the results and contributed to the final manuscript.
3. **R. Schneider**, T. Oberhauser, T. Klatt, F. A. Gers, and A. Löser, "RelVis: Benchmarking OpenIE Systems," presented at the International Semantic Web Conference (Posters, Demos & Industry Tracks), 2017.
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework.
   - R. Schneider led the verification of the analytical methods with input from F. A. Gers, and A. Löser.
   - R. Schneider led the implementation of the theory, baselines, and execution of computations and was assisted by T. Oberhauser and T. Klatt.
   - R. Schneider was steering various in-depth discussions with F. A. Gers, and A. Löser that led to the final experiments and results.
   - All authors discussed the results and contributed to the final manuscript.

4. **R. Schneider**, S. Arnold, T. Oberhauser, T. Klatt, T. Steffek, and A. Löser, "Smart-MD: Neural Retrieval of Medical Topics," in Companion of The Web Conference 2018 on The Web Conference 2018, 2018, pp. 203–206.
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework with support from S. Arnold.
   - R. Schneider led the verification of the analytical methods with input from S. Arnold, and A. Löser.
   - R. Schneider led the implementation of the theory, baselines, and execution of computations and was assisted by T. Oberhauser, T. Klatt and T. Steffek.
   - R. Schneider was steering various in-depth discussions with S. Arnold, and A. Löser that led to the final experiments and results.
   - All authors discussed the results and contributed to the final manuscript.
5. S. Arnold, **R. Schneider**, P. Cudré-Mauroux, F. A. Gers, and A. Löser, "SECTOR: A Neural Model for Coherent Topic Segmentation and Classification," TACL, vol. 7, pp. 169–184, 2019
   Author Contributions:
   - S. Arnold conceived the presented idea.
   - R. Schneider and S. Arnold designed the theoretical and computational framework and model architecture of SECTOR.
   - R. Schneider and S. Arnold led the verification of the analytical methods with input from F. A. Gers, A. Löser and P. Cudré-Mauroux.
   - R. Schneider and S. Arnold implemented the theory, baselines, performed the computations and analysed the data.
   - R. Schneider and S. Arnold were steering various in-depth discussions with F. A. Gers and A. Löser that led to the final experiments, model design, dataset and results.
   - All authors discussed the results and contributed to the final manuscript.

6. **R. Schneider**, T. Oberhauser, P. Grundmann, F. A. Gers, A. Löser, and S. Staab, "Is language modeling enough? Evaluating effective embedding combinations," in Proceedings of the 12th language resources and evaluation conference, Marseille, France, May 2020, vol. 12, pp. 4741–4750.
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework.
   - R. Schneider led the verification of the analytical methods with input from F. A. Gers, A. Löser and S. Staab.
   - R. Schneider led the implementation of the theory, baselines, and execution of computations and was assisted by T. Oberhauser and P. Grundmann.
   - R. Schneider was steering various in-depth discussions with F. A. Gers, A. Löser and S. Staab that led to the final experiments, model design, dataset and results.
   - All authors discussed the results and contributed to the final manuscript.
7. **R. Schneider**, M. Mayrdorfer, H. Schmidt, K. Budde, F. A. Gers, A. Löser, and S. Staab, “SmartMD: Deep Learning enabled Differential Diagnosis”
   Author Contributions:
   - R. Schneider conceived the presented idea and was the main contributor.
   - R. Schneider designed the theoretical and computational framework with support from M. Mayrdorfer.
   - M. Mayrdorfer directed the verification of the medical soundness. H. Schmidt assisted with this task.
   - R. Schneider led the verification of the analytical methods with input from M. Mayrdorfer, K. Budde, F. A. Gers, A. Löser, and S. Staab.
   - R. Schneider and M. Mayrdorfer supervised and conducted the experiments and interviews with medical doctors.
   - R. Schneider led the implementation of the theory and baselines, performed the computations and analysed the data.
   - R. Schneider was steering various in-depth discussions with M. Mayrdorfer, K. Budde, F. A. Gers, A. Löser and S. Staab that led to the final experiments, framework design and results.
   - All authors discussed the results and contributed to the final manuscript.

Contents

List of Abbreviations

ADAM	Adaptive Moment Estimation - A Method for Efficient Stochastic Optimization.
BE / BERT	Bidirectional Encoder Representations from Transformers.
BLSTM	Bidirectional Long Short-term Memory (Neural Network Architecture).
CBOW	Continuous Bag of Words.
CDS	Clinical Decision Support.
CDSS	Clinical Decision Support System.
CIE	ClausIE Open Information Extraction System.
CNN	Convolutional Neural Network.
CPT	Current Procedural Terminology.
CR	Customer Review Sentiment Analysis Task in the SentEval Benchmark.
CUI	Concept Unique Identifiers (UMLS).
DDx	Differential Diagnosis.
DL	Deep Learning.
e.g.	Exemplum Gratia (English: for example).
EHR	Electronic Health Record.

EL	ELMO (Language Model Architecture).
ELMO	Embeddings from Language Model.
Eq.	Equation.
FT	fastText.
GAN	Generative Adversarial Networks.
GPT	Generative Pre-trained Transformer.
GPU	Graphics Processing Unit.
ICD	International Statistical Classification of Diseases and Related Health Problems.
ICU	Intensive Care Unit.
IDF	Inverse Document Frequency.
IE	Information Extraction.
INDREX	In-Database Relation Extraction System.
IR	Information Retrieval.
LDA	Latent Dirichlet Allocation.
LOINC	Logical Observation Identifiers Names and Codes.
LSTM	Long Short-Term Memory (Neural Network Architecture).
MAP	Mean Average Precision.
MIMIC-III	Medical Information Mart for Intensive Care (Dataset).
MPQA	Opinion-Polarity Task in the SentEval Benchmark.
MR	Moview Review Sentiment Analysis Task in the SentEval Benchmark.
MRI	Magnetic Resonance Imaging.
MRPC	Paraphrase Detection Task in the SentEval Benchmark.
NEL	Named Entity Linking.

NER	Named Entity Recognition.
NLP	Natural Language Processing.
NPMI	Normalized Point-wise Mutual Information.
NYT-222	Open Information Extraction Dataset Proposed by Mesquita, Schmidek, and Barbosa, 2013.
OIE	Open Information Extraction.
OIE-2016	Open Information Extraction Dataset Proposed by Stanovsky and Dagan, 2016.
PA	Entity Embedding Proposed by Pappu et al., 2017.
PENN-100	Open Information Extraction Dataset Proposed by Y. Xu et al., 2013.
PMI	Pointwise Mutual Information.
POS	Part-of-speech Tagging.
PP	PredPat Open Information Extraction System.
PubS	PubMedSection Dataset.
RDBMS	Relational Database Management System.
RQ	Research Question.
SICK	Sentences Involving Compositional Knowledge - Textual Similarity Dataset.
SICK-E	Natural Language Inference Classification Task in the SentEval Benchmark).
SIE	Stanford Open Information Extraction System.
SOTA	State of the Art.
SP	SECTOR Pubmed. SECTOR Model Trained on PubMed Articles.
SQL	Structured Query Language.
SST	Sentiment Analysis Task in the SentEval Benchmark based on the stanford Sentiment Treebank.
SUBJ	Subjectivity Status Task in the SentEval Benchmark.

SW	SECTOR Wikipedia. SECTOR Model Trained on Wikipedia Articles.
TASTY	Tag-As-You-Type a Text Editor for Interactive Entity Linking.
TF	Term Frequency.
TF-IDF	Term Frequency Inverse Document Frequency.
TREC	Question Type Classification Task in the SentEval Benchmark.
UMLS	Unified Medical Language System.
URL	Uniform Resource Locator.
WEB-500	Open Information Extraction Dataset Proposed by Mesquita, Schmidek, and Barbosa, 2013.
WikS	WikiSection Dataset.

Zusammenfassung

Bei der Patientenversorgung verfasst medizinisches Fachpersonal große Mengen klinischer Dokumente. Diese dokumentieren medizinische Fälle von der Anamnese bis zum jeweiligen klinischen Endpunkt. Das automatisierte Analysieren und Finden relevanter Krankenakten bietet die Möglichkeit, Ärzte in großem Umfang bei ihrer täglichen Arbeit zu unterstützen. Das automatisierte Verständnis von klinischem Text ist jedoch nicht trivial, insbesondere das Verarbeiten von Pflegenotizen und Diagnoseberichten stellt eine Herausforderung dar. Klinische Dokumente weisen eine hohe Varianz in Länge, Struktur, Vokabular sowie lexikalischer und grammatikalischer Korrektheit auf. Häufig sind sie stark vom jeweiligen klinischen Kontext abhängig. Aus diesen Gründen scheitern Ansätze, die auf syntaktischen Regeln und diskreter Textrepräsentationen basieren, oft.

Diese Arbeit befasst sich mit dem Entwurf und der Evaluierung von Methoden und Modellen, die sowohl generalisierbar als auch anpassungsfähig genug sind, um klinische Texte automatisch zu analysieren. Ziel dieser Arbeit ist es, die Grundlagen textbasierter klinischer Entscheidungsunterstützungssysteme zu verbessern. Textbasierte klinische Entscheidungsunterstützungssysteme können das Wissen in Krankenhausarchiven und medizinischen Publikationen im Alltag von Ärzten nutzbar machen. Solche Systeme müssen der wachsenden Menge an klinischen Dokumenten in Krankenhausarchiven gerecht werden. Ein Kernproblem für textbasierte klinische Entscheidungsunterstützungssysteme besteht in der ganzheitlichen Repräsentation von Patientendaten für die automatisierte Verarbeitung. Wir begegnen diesen Herausforderungen, indem wir ein Framework für die Deep-Learning-basierte Differenzialdiagnoseunterstützung entwerfen.

Unter Betrachtung der genannten Anforderungen entwerfen und evaluieren wir Methoden zur automatischen Analyse von medizinischen Texten, basierend auf drei Informationsrepräsentationsparadigmen: (1) Diskrete Relationsextraktion unter Verwendung des “Open Information Extraction” Paradigmas. (2) Neuronale Textrepräsentationen basierend auf Sprach- und Themenmodellierung. (3) Kombinieren komplementärer neuronaler Textrepräsentationen.

Unser Framework übersetzt klinische Diagnoseschritte und Pfade in statistische und Deep-Learning-basierte Modelle. Wir zeigen, dass Deep-Learning basierte Differenzialdiagnosesysteme von kontextualisierten Sprachmodellen profitieren. In einem umfassenden Benchmark identifizieren wir Defizite des “Open Information Extraction” Paradigmas welche eine Anwendung auf klinische Texte erschweren. Wir entwerfen ein kontextualisiertes Textrepräsentationsmodell basierend auf Themenmodellierung. Unsere Ergebnisse zeigen, dass neuronale Textrepräsentationen welche auf Themenmodellierung basieren, Informationen abbilden welche komplementär sind, zu auf Sprachmodellierung basierenden Ansätzen. Unsere Experimente mit Ärzten, basierend auf einer prototypischen Implementierung, validieren den Deep-Learning-gestützten Differentialdiagnoseprozess. Darüber hinaus identifizieren wir auf Basis unserer qualitativen und quantitativen Erkenntnisse sieben Designherausforderungen für textbasierte klinische Entscheidungsunterstützungssysteme.

Abstract

Medical professionals create vast amounts of clinical texts during patient care. Often, these documents describe medical cases from anamnesis to the final clinical outcome. Automated understanding and selection of relevant medical records pose an opportunity to assist medical doctors in their day-to-day work on a large scale. However, clinical text understanding is challenging, especially when dealing with clinical narratives such as nursing notes or diagnostic reports. These clinical documents differ extensively in length, structure, vocabulary, and lexical and grammatical correctness. In addition, they are highly context-dependent. For all these reasons, approaches based on syntactic rules and discrete text representation often fail to address the variety of clinical narratives propagating unrecoverable errors to downstream applications.

Therefore, this thesis focuses on evaluating and designing methods and models that are generalizable and adaptable enough to deal with these challenges. Our goal is to enable text-based clinical decision support systems to utilize the knowledge from clinical archives and medical publications. We aim to design methods that can scale up to the growing amount of clinical documents in hospital archives. A fundamental problem in achieving deep-learning-enabled clinical decision support systems is designing a patient representation that captures all relevant information for automated processing. We engage these challenges by designing a framework for deep-learning-enabled differential diagnosis support. Guided by the needs emerging from this framework, we design and evaluate methods based on three information representation paradigms: (1) Discrete relation extraction using the open information extraction paradigm. (2) Neural text representations

based on language and topic modeling. (3) Combining complementary neural text representations.

Our framework translates clinical diagnostic steps and pathways to statistical and deep-learning-based models. Accordingly, we can show that deep-learning-enabled differential diagnosis benefits from contextualized information representations. Further, we identify shortcomings of the open information extraction paradigm in a comprehensive benchmark. We design a distributed text representation model based on topical information. Our extensive large-scale experiment results show that topical distributed text representations capture information complementary to language modeling-based approaches across domains, thus enabling a holistic text representation for medical texts. Our experiments with medical doctors using our prototypical implementation of the deep-learning-enabled differential diagnosis process validate this framework. Moreover, we identify seven crucial design challenges for text-based clinical decision support systems based on our qualitative and quantitative findings.

Chapter 1

Introduction

Medical professionals create vast amounts of clinical texts during patient care. Often, these documents describe medical cases from anamnesis to the final clinical outcome (Shickel et al., 2018). Accordingly, clinical archives pose a valuable source of knowledge when dealing with uncommon complications and rare diseases (Faviez et al., 2020; Ronicke et al., 2019; Shen and H. Liu, 2018). This knowledge is often unused due to its inaccessibility to medical practitioners.

Automated understanding and selecting relevant medical records pose the opportunity to solve this issue (Demner-Fushman, Chapman, and McDonald, 2009). Moreover, it enables automated *Clinical decision support systems (CDSS)* to support medical practitioners in information-seeking and decision processes. CDSS can serve a wide variety of tasks, such as Patient Safety (Eslami et al., 2012; Mahoney et al., 2007; McEvoy et al., 2017), Clinical Management (McMullin et al., 2004 Sep-Oct; Salem et al., 2018), Diagnostic Support (Cui, Bozorgi, et al., 2012; De Fauw et al., 2018; Ronicke et al., 2019). Many diagnostic support systems aim at a small subfield of medical care and often do not exploit data collected in EHRs (De Fauw et al., 2018; Goldenberg, Nir, and Salcudean, 2019; D. Jiang et al., 2020; Y. Liu et al., 2019).

However, clinical text understanding is a challenge, especially when dealing with clinical narratives such as nursing notes or diagnostic reports. These clinical documents differ extensively in length, structure, vocabulary, and lexical and grammatical correctness. For all these reasons, approaches based on syntactic

rules and discrete text representation often fail to address the variety of clinical narratives propagating unrecoverable errors downstream (Hong et al., 2018; Leaman, Khare, and Lu, 2015; Pink, Nothman, and Curran, 2014; Starlinger et al., 2017).

Therefore, this thesis focuses on evaluating and designing methods and models that enable text-based clinical decision support systems to utilize the knowledge from clinical archives and medical publications.

1.1. Clinical Information Management

Documenting the state and progress of a patient is a critical task in hospitals and clinics. Therefore, medical practitioners store structured information such as lab results and screening scores or unstructured data such as nursing notes, doctors' letters, or radiology reports in *Electronic Health Records (EHR)*. These records document the patients' trajectory and taken clinical pathways, including their clinical endpoints (Shickel et al., 2018). The first entries in an EHR contain the anamnesis and an initial evaluation of the patient's trajectory. This initial trajectory describes the severity of the case and the most likely outcomes. Correspondingly, the medical professional initiates the most appropriate clinical pathway consisting of a line of diagnostics and treatments (Z. Huang, Dong, et al., 2014). The medical staff updates the EHR regularly during the whole process. These updates include nursing notes and diagnostics results, such as radiologic images and the corresponding radiology reports. The nursing notes and diagnostics reports are usually written as free-form text to provide the crucial details and context of the patients' situation. Moreover, continually taken structured measurement data like body temperature or oxygen saturation might also be included in these updates. The medical personnel is reevaluating the patients' current trajectory constantly and adapts the clinical pathway as needed. Finally, the patients arrive at a clinical endpoint, such as being cured. Conclusively, EHRs are a valuable and dense source of medical knowledge. Therefore, gaining new medical insights and improving treatment quality based on EHRs is an important goal (Aspland, Gartner, and Harper, 2021; Landi et al., 2020).

Medical doctors see many patients with common and rare symptoms, diseases, or complications. In particular rare cases are challenging for medical doctors,

especially if they are at the beginning of their career and cannot rely on many years of experience. To deal with situations of uncertainty, medical practitioners can consult colleagues, medical guidelines, literature, or the clinics' archives for similar cases. Medical doctors often tend to stick to the first three options. As a result, practitioners are not utilizing information collected in EHRs (Fu et al., 2020).

The most significant problem is the inaccessibility of information stored in clinical archives (Fu et al., 2020; R. T. Sutton et al., 2020). EHRs already pose the opportunity to access clinical archives effectively, but the ever-growing amount of medical notes in EHRs is overwhelmingly large. For example Charité Berlin handled 806.524 cases in 2021[1]. Modern EHR management systems such as T-Base (Schmidt et al., 2021) help clinicians retrieve EHRs but identifying critical clinical concepts of each case and determining its relevance requires medical practitioners to read the free-form text. This situation poses an immediate problem since medical professionals often have minimal time boxes per patient and experience fast-paced work intensification (Huhtala et al., 2021). Furthermore, even if a doctor collects similar cases to the case at hand, she still needs to group, limit, and filter those selected patients. Finally, she needs to evaluate and rank the clinical pathways taken in the past for suitability, which is also a complex and time-consuming task. Consequently, the doctor may miss examining clinical pathways that were proven effective in the past (van der Vegt et al., 2020). Worse, this might also hinder the research and development of new therapies or diagnostic protocols.

1.2. Supporting Clinicians with Text-Based Decision Support Systems

One way of solving the problem is to apply *natural language processing (NLP)*, *computer vision*, and *information retrieval (IR)* methods. Using methods from the fields of information retrieval enables operationalizing the information needs of medical professionals (Ely et al., 2000). Computer vision and natural language processing can extract information hidden in unstructured medical imaging and text data (Fu et al., 2020; R. T. Sutton et al., 2020). These three approaches

[1] `https://www.charite.de/die_charite/profil/zahlen_fakten/`

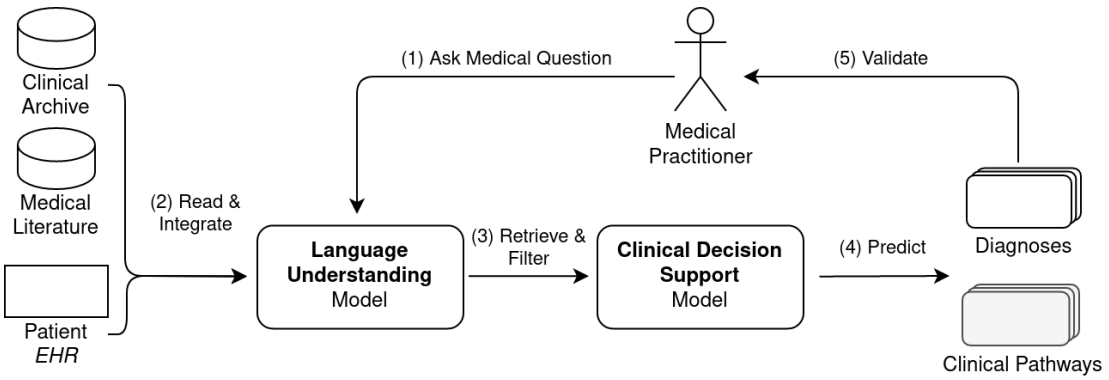


Figure 1.1.: The text-based differential diagnosis support process.

enable designing methods that can considerably increase the accessibility of clinical archives for practitioners' day-to-day use cases. However, every part of this solution poses complex challenges and requires well-evaluated design decisions.

One practical approach is to frame the text-based clinical decision support process as a recommendation task to suggest complementary clinical pathways or single steps to medical practitioners based on observations made in the clinical archive (see Figure 1.1). NLP methods such as *Open Information Extraction* (Banko et al., 2007; Q. Wei et al., 2020), *Clinical Concept Recognition* (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017), and *Topic Modeling* (Blei, 2012) can derive these observations from free-text documents (2). Medical computer vision models can provide additional clinical concepts (Esteva et al., 2021). A holistic model of the patients' current trajectory enables the IR methods to retrieve, group, and filter similar patients (H. Liu et al., 2013 -3- 18) (3). Based on this cohort of relevant patients, it is possible to recommend complementary clinical pathways to medical practitioners (4). Finally, the medical staff validated the recommendations, giving feedback to the system (5). Although this solution seems straightforward, several challenges to consider hinder naive approaches from being successful.

High Context-Dependence. First, medical data is highly context-dependent (Sharafoddini, Dubin, and J. Lee, 2017; Starlinger et al., 2017) and requires approaches incorporating the patients' current situation as completely as possible. Methods with this capability are, for example, *Deep-Learning-based (DL)* models (Esteva et al., 2021; M. E. Peters et al., 2018). However, these models require a

vast amount of training data to be successful (J. Lee, Yoon, et al., 2019; Radford, J. Wu, Child, et al., 2019; Vaswani et al., 2017).

Limited Access to Labeled Data. Second, there is a lack of available data due to the high data protection standards. Available datasets are rather small and focus on medical subfields such as intensive care (Johnson et al., 2016) or oncology (Borchert et al., 2020) This problem affects quality measurements, training data for supervised and unsupervised machine learning models. This circumstance is even more problematic in research since it hinders the comparability and reproducibility of studies conducted by different researchers using exclusive clinic specific data sets.

Large Quantities of Data. Third, the amount of data in a typical hospital is considerable (Schmidt et al., 2021) and requires infrastructure and methods to handle large volumes of data with low response times.

High Variance, Ambiguity, and Noise. Fourth, models in clinical settings need to perform with precision while being able to generalize to unseen data. Generalization capabilities are essential for NLP models (Shickel et al., 2018). Medical professionals often write clinical notes while interacting with patients utilizing department- and context-specific abbreviations, writing styles and highly specialized terms. Accordingly, clinical notes often contain spelling errors, staccato-like sentence fragments, and grammatical errors (Starlinger et al., 2017; van Aken, Trajanovska, et al., 2021). These irregularities add additional challenges to the already present challenge of ambiguity of medical concepts and their abbreviations.

General Multi-Modal Patient Representation. Fith, to enable high-quality recommendations, it is imperative to design a patient representation independent of the medical profession. This representation needs to cover the problem space's multi-modality and be beneficial for information retrieval methods (Shickel et al., 2018; R. T. Sutton et al., 2020).

This thesis approaches the problem by designing a framework for designing differential diagnosis (DDx) support systems that utilize data stored in EHRs. Due

to the complexity of the problem, we discuss exclusively text-based methods for clinical decision support in this thesis.

1.3. Clinical Text Understanding

Written clinical narratives highly depend on context and contain dense information about patients (Starlinger et al., 2017). These documents are the primary way to communicate between hospital medical teams (Shickel et al., 2018).

The clinical text understanding component is an integral part of a *text-based clinical decision support system (CDSS)*. The common goal of clinical text understanding models is to extract information from the free-form text to be available for clinical decision support systems (Shickel et al., 2018; R. T. Sutton et al., 2020). The resulting text representation can be a valuable component of patient representations. Choosing knowledge to represent and representation methods influence CDSS to a large extent. Classical approaches often use discrete feature vectors obtained with information extraction models (Cui, Bozorgi, et al., 2012; L. Li et al., 2015; Sarmiento and Dernoncourt, 2016). The recent success of latent and distributed neural text representations opens new possibilities for clinical use-cases and patient representations (Glicksberg, Miotto, et al., 2018; Gu et al., 2021).

1.3.1. Discrete Representations

A central task in clinical text understanding is recognizing and disambiguating bio-medical concepts mentioned in a text. Classical approaches often solve this task by applying syntactic extraction patterns (Aronson and Lang, 2010 May-Jun; Tseytlin et al., 2016) to identify a clinical concept's beginning and end position in a text. As a next step, a model disambiguates the mention against a controlled vocabulary or knowledge base that contains all clinical concepts of interest (Fu et al., 2020). *Text segmentation* (Beeferman, Berger, and Lafferty, 1999) and *topic modeling* (Blei, 2012; L. Wang, S. Li, et al., 2017) pose additional signals for such models by dividing texts into coherent sequences of sentences.

Clinical Concept Recognition models aim to extract mostly sentence-local information, but the interaction between multiple clinical concepts can span multiple

sentences or paragraphs. Therefore, *relation extraction models* may take sentence-spanning context into account. Classical approaches to relation extraction use syntactic extraction patterns to extract relations according to an a priory specified schema (Chiticariu, Krishnamurthy, et al., 2010; Kilias, Löser, and Andritsos, 2015).

Open Information Extraction (OIE) aims to extract relations between entities or clinical concepts from large text corpora without pre-specifying a target schema (Banko et al., 2007). An OIE system could extract all mentioned drug-drug interactions from a text without specifying the number of involved drugs, the relation type or schema, a priori (Nebot and Berlanga, 2012). For example, an OIE system might extract the following 3-ary relation from a medical research paper (Triplitt, 2006): (`Metformin, cimetidine,` *can compete for*`, elimination`)[1]. Most OIE methods depend on intermediate syntactic representations such as dependency trees (C. Manning and Schutze, 1999; Nivre et al., 2016) or part-of-speech-tags (Jurafsky and James, 2021; C. Manning and Schutze, 1999; Marcus, Santorini, and Marcinkiewicz, 1993).

Clinical narratives often contain modifiers regarding the mentioned clinical concepts, such as negations. For downstream applications, *assertion detection* models must extract these modifiers correctly (van Aken, Trajanovska, et al., 2021).

Finally, the extracted information can populate databases, knowledge graphs, or utilize decision support systems. This classical view of the clinical text understanding models often leads to discrete patient representations e.g., a bag of clinical concepts represented as a one-hot vector (Cui, Bozorgi, et al., 2012; Soysal et al., 2018).

1.3.2. Latent Representations

With the reemergence of neural networks for solving natural language processing tasks and their vast success, many authors propose deep learning models for clinical text understanding (Esteva et al., 2021; J. Lee, Yoon, et al., 2019). Deep Learning enables models to generalize to a wide variety of unseen data. Moreover, it is

[1]This example was produced using Stanford core NLP 4.2.2:
`https://stanfordnlp.github.io/CoreNLP/history.html`

possible to combine multiple tasks in one model and profit from complementary error signals, propagating back through specialized layers (Radford, J. Wu, Child, et al., 2019; Raffel et al., 2020). Furthermore, neural networks enable pretraining on basic unsupervised language understanding tasks and reusing the pre-trained models for supervised training while significantly improving their performance on the supervised task (Devlin et al., 2019). The reason for this property is that deep neural networks can learn effective latent representations of text. These neural text representations can contribute to a latent patient representation for clinical decision support tasks (Glicksberg, Miotto, et al., 2018; Gu et al., 2021; Landi et al., 2020; Miotto, L. Li, et al., 2016).

1.4. The Scope of this Thesis

This thesis focuses on designing a generally applicable CDSS framework and investigating text representation methods that benefit such systems. We approach the challenges of clinical decision support by proposing a framework and medical text understanding models to address the three subproblems: *Open Information Extraction*, *Medical Text Representation*, and *Clinical Decision Support*. We evaluate the discrete open information extraction paradigm for its suitability in clinical settings. Furthermore, we use the deep learning paradigm to explore distributed latent text representations. Deep learning-based methods enable us to combine hidden probability features learned from a wide variety of unsupervised tasks and datasets. Moreover, methods based on deep learning are language-agnostic and can be further specialized on clinic-specific vocabularies. In addition, we address the challenges of high context dependence, training data efficiency, runtime efficiency, and robustness regarding variance, ambiguity, and noise in clinical text data. Throughout this thesis, we focus on evaluating and designing language understanding models that are efficient in the clinical domain. We provide results and insights from a preliminary user study.

1.4.1. Research Objectives

This thesis addresses the problems of automatic clinical decision support and clinical text understanding. The central hypothesis of this thesis is the following:

The differential diagnosis process can be the foundation for developing deep-learning-enabled text-based clinical decision support systems. These systems rely extensively on text and patient representations, and integrating discrete and latent representations enhances their ability to construct a comprehensive patient model. Complementary combinations increase the accuracy and helpfulness of the system. We divide this general problem into the following four research questions (RQ):

RQ1: Is the Open Information Extraction Paradigm Suitable for Clinical Text Understanding? Extracting relations between entities such as clinical concepts is a central task in Information Extraction. Classical approaches require specifying relations a priory and often rely on specifically designed extraction rules. The ever-growing number of medical documents and the Zipfian nature of text prohibits expensive human-curated adaption of these rules. The Open Information Extraction paradigm aims to resolve the requirement of a priory specified relation schemata. Nevertheless, when comparing the extraction result mentioned earlier, "`(Metformin, cimetidine,` *can compete for*`, elimination)`", with the originating sentence by Triplitt, 2006: "Metformin and cimetidine, both cationic (positively charged) drugs, can compete for elimination through kidneys by renal tubular secretion." it becomes clear that OIE systems might lose important context information. The dependence of most OIE methods on intermediate syntactic representations is an additional challenge when confronted with clinical narratives (Starlinger et al., 2017).

Text representations for CDSS require that rare entities and their relations are extracted with high recall and without the loss of context. Clinical concepts and assertive modifiers must be recognized accurately. As a result, we aim to investigate the applicability of OIE systems in text-based clinical decision support systems.

RQ2: Can Neural Text Representations aid Text-based Clinical Decision Support Systems? Neural Text Understanding Models commonly rely on distributed text representations learned using pretraining tasks on large, diverse datasets often taken from the web. The most common approach to obtain such text representations is to train deep neural networks on unsupervised language generation tasks. These neural text representations are helpful in downstream applications

to capture the semantics of words in their context. By the nature of the widely used language modeling pretraining task, these representations focus on local context. A crucial capability is to capture the global context describing a patient's situation. Therefore, a clinical text representation must be able to identify coherent passages in medical texts, recognize their topical facet, and capture single sentences' meaning given the context of the entire document.

Which type of representation is best suited to provide knowledge from EHRs and medical literature, capturing both local and global context, is unclear.

RQ3: Are Text Representations Trained with Differing Pretraining Goals Complementary? Language Modeling is a frequent pretraining task for text representations used as a foundation in transfer learning settings. Other pretraining tasks, such as topic modeling or entity linking, require the model to focus on different aspects of a text. Therefore, they capture varying granularity, context size, and modality information. It is mandatory for text understanding models to capture the meaning of clinical narratives and medical research text as completely as possible to achieve the best result as a part of a patient representation. Identifying and combining complementary text representations can lead to holistic representations that improve text-based clinical decision support systems. Whether combining multiple generic and specialized text representations benefits CDSS is an open question.

RQ4: How Effective are Deep Learning Enhanced Medical Information Seeking Processes? The differential diagnosis process supports medical practitioners in finding the most effective clinical pathway. We expect this process to be enhanced further by adding additional signals collected from hospital archives and medical literature. Therefore, we design a Deep Learning enabled differential diagnosis framework. This framework allows employing Deep Learning models for typical tasks in the process, assisting medical practitioners. A prototypical implementation of this framework should receive positive feedback when used by medical practitioners.

Medical literature is another valuable source for supporting doctors in their day-to-day work. Doctors might consult medical literature databases such as

PubMed[1] when searching for additional information regarding a complicated case. However, even with PubMed's filter capabilities, sorting, selecting, and skimming the relevant literature is still time-consuming (Vibert et al., 2009; Yoo and Mosa, 2015). A paragraph answer retrieval system that selects and ranks topically coherent passages from medical literature could reduce the time spent on literature research (Sarrouti and Ouatik El Alaoui, 2017). Such a system requires a model to identify semantically coherent and relevant passages. Typical tasks of neural topic models involve segmenting texts and assigning topical labels to those passages. The intermediate distributed text representation such models create in their hidden layers must represent this information to solve segmentation and topic classification tasks. Therefore, this intermediate representation might pre-cluster texts and adequately represent passage retrieval.

1.4.2. Contributions

The main contribution of this thesis is the design and application of the deep-learning-enabled differential diagnosis process. This thesis will focus on three main aspects of text-based clinical decision support systems: first discrete information extraction and representation; second neural information representation; third, the design and application of the deep-learning-enabled differential diagnosis process. We investigate these aspects with respect to the research objectives stated in Section 1.4.1. We summarize our work accordingly and provide insights into our theoretical, practical, and empirical contributions:

Analyzing discrete information extraction and representation based on OIE

- We analyze OIE systems on multiple datasets and reveal a lack of stringent task formulation and annotation policies. We observe that syntactic taggers are a frequent error source that propagates errors down to the OIE system. (Section 3.2.6)
- We find that the analyzed OIE systems often extract unnormalized and over-specific relation tuples. (Section 3.2.6)

[1] `https://pubmed.ncbi.nlm.nih.gov/`

- We argue that the reviewed OIE systems, which are already overfitting and struggling with news datasets, are not suitable for application in clinical narratives. (Section 3.2.6)
- We present RelVis, the first integrated benchmarking system for quantitative and qualitative evaluation of OIE systems. (Section 3.2.5)
- We present a novel OIE system based on an in-memory database and report execution times on large datasets in seconds and easy integration with relational data. (Section 3.1.4 & Section 3.1.5)

Neural Text Representations for Clinical Applications

- We introduce PubMedSection, a novel dataset for medical topic segmentation and classification. (Section 4.3)
- We extend the "senteval" benchmark (Conneau and Kiela, 2018) with WikiSection (Arnold, Schneider, et al., 2019) and PubMedSection. (Section 4.4.2)
- We compare specialized text embeddings with general-purpose embeddings. We report that language models lack topical information. (Section 4.4)
- We identify effective embedding combinations that yield holistic text representations and achieve new state-of-the-art results in senteval. (Section 4.4.3)

Deep Learning enabled Clinical Decision Support

- We demonstrate that neural clinical concept recognition and topic segmentation enable clinicians to search for topical facets of diseases. Furthermore, we show that neural topic models such as SECTOR (Arnold, Schneider, et al., 2019) allow selecting relevant answer passages. (Section 5.1.2)
- We propose the deep-learning enabled differential diagnosis process as a framework to formalize and implement differential diagnosis support systems. (Section 5.2.2)

- We present results from our qualitative and observation studies involving five clinicians that validate the deep-learning enabled differential diagnosis process. (Section 5.2.4 & Section 5.2.5.1)
- We identify seven design challenges crucial for research and practical application of text-based clinical decision support systems. (Section 5.2.5.2)

In addition to this thesis's main contributions, we provided open-source implementations of our work and contributed to open-source frameworks such as OpenNLP[1] and DeepLearning4j[2]. In this context, we have supervised theses that explore additional lines of thought, recreate experiments from related literature, and verify our work in practical applications. See Appendix A for details.

1.4.3. Limitations

Some approaches to designing text-based CDSS focus on specific auxiliary tasks, such as clinical concept linking, assertion detection, or medical question answering. Many CDSS are specialized, for example, on the properties of a specific disease. In this thesis, we do not address the challenges that arise when focussing on a specific disease or auxiliary task.

We do not discuss every aspect of the clinical decision-support process, as shown in figure 1.1. We limit our scope to text data in this thesis and do not discuss the problem of integrating latent text, image, and time-series representations into a combined multi-modal patient representation. Also, we use for the candidate retrieval (3) and clinical pathway prediction step (4) baseline models. Instead, we focus on evaluating and designing clinical language understanding models that are efficient in the clinical domain. Moreover, a clinical study on the effectiveness of our proposed framework, extending the results of our preliminary user study, is beyond the scope of this thesis.

1.5. Thesis Outline

We structure this thesis along the vision of the Deep Learning enabled differential diagnosis process and three main topics exploring text representations for text-

[1] `https://opennlp.apache.org/`
[2] `https://deeplearning4j.konduit.ai/`

based clinical decision support systems.

Chapter 1 - Introduction. We introduce the vision of the Deep Learning enabled differential diagnosis process and motivate the need for robust, efficient text-based clinical decision support systems. We discuss the challenges in medical text understanding. We divide the main problem into three topics and motivate five related research questions, which we explore in this thesis.

Chapter 2 - Background. We discuss the existing literature regarding text-based clinical decision support and medical text understanding. We review the concept of discrete text representation using features extracted with OIE methods. We further summarize the idea of latent distributed language representations, which pose a fundamental groundwork for Deep Learning based language understanding models. Moreover, we relate the idea of neural text representations with patient representation models based on related literature.

Chapter 3 - Analysing Open Information Extraction. Text-based CDSS require high robustness, recall and, fast execution times from OIE systems to be useful. Therefore, we create a benchmark to compare approaches to Open Information Extractions. We analyze errors and complete error classes described in the literature with our results. We design RelVis, a benchmarking and error analysis tool for open information extraction systems, to perform this analysis. Additionally, we explore integrating OIE methods in a main-memory database system to increase execution performance.

Chapter 4 - Neural Text Representations for Clinical Applications. We explore neural network-based latent text representations for medical use-cases. We observe that medical literature, as well as clinical narratives, are structured with topical coherent passages. Identifying locally dominant topics and their boundaries results in a valuable signal for medical text understanding models. We engaged this challenge in-depth in Arnold, Schneider, et al., 2019. We analyze and benchmark the resulting latent neural text representation and compare it to other specialized and general purpose text representations. Moreover, we show

that our model encodes knowledge that language modeling-based approaches, such as ELMo (M. E. Peters et al., 2018) and BERT (Devlin et al., 2019), miss.

Chapter 5 - Deep Learning enabled Clinical Decision Support. We design the Deep Learning enabled differential diagnosis support process to assist medical practitioners in their day-to-day work. We analyze the steps necessary and model each step as problems solvable with statistical and deep learning methods. We identify medical passage retrieval and clinical pathway recommendation as core problems. We solve both challenges in prototypes building upon our analyzes and models. Furthermore, we validate the Deep Learning enabled differential diagnosis process with a user study conducted on the clinical pathway recommendation prototype.

Chapter 6 - Conclusion and Future Work. Finally, we conclude this thesis. We discuss our results regarding the research objective formulated in section 1.4.1. In addition, we present possible business perspectives that arise from the Deep Learning enabled differential diagnosis support process.

Chapter 2

Background

In this chapter, we discuss the theoretical background of this thesis. First, we discuss the field of computer-aided clinical decision support in Section 2.1. We review text-based clinical decision support methods and common tasks such as clinical concept recognition, clinical concept linking, cohort identification, and clinical pathway recommendation. In section 2.2, we introduce the concept of open information extraction. Following this, we discuss in section 2.3 the foundations of distributed language representations, such as the Vector Space Model, language models, topic models, and neural word embeddings. Finally, we discuss in section 2.4 how this thesis relates to the aforementioned related literature.

2.1. Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) intend to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge, patient information, and other health information (R. T. Sutton et al., 2020). Traditional CDSS gather patient-specific characteristics and match them to a medical knowledge base. Based on that, the system recommends clinical actions, diagnoses, or clinical pathways to the clinician. Typical tasks of CDSS are disease-specific diagnostic support, such as cancer detection on medical imaging, patient trajectory modeling, cohort selection, and general-purpose clinical outcome prediction. Computer-aided CDSS have been in the focus of research for multiple decades until now.

Electronic Health Records (EHR) have been adopted widely (84% of US hospitals in 2018) (Shickel et al., 2018) and pose a new clinical knowledge source for decision support systems. EHR systems store data associated with each patient encounter. On the one hand, this data includes structured data such as demographic information, diagnoses, performed laboratory tests and results, and prescriptions. On the other hand, from an information systems perspective, unstructured data such as medical images, diagnostics reports, clinical notes, and other clinical narratives. The free-form text reports, notes, and letters are essential to communication between medical practitioners. They provide detailed information about a patient's situation and crucial context to each EHR (Shickel et al., 2018; R. T. Sutton et al., 2020).

The primary purpose of EHRs is internal hospital tasks such as administration, billing, archiving medical data, and communication between medical practitioners. Therefore, the data is often linked with medical classification schemas, controlled vocabularies, and knowledge bases. Some examples include diagnosis codes such as the International Statistical Classification of Diseases and Related Health Problems (ICD), procedure codes such as the Current Procedural Terminology (CPT), laboratory observations such as the Logical Observation Identifiers Names and Codes (LOINC), and medication codes such as RxNorm (Shickel et al., 2018). EHRs at hospitals and clinics can improve patient care by minimizing errors, increasing efficiency, and improving care coordination while providing a rich data source for researchers (Knake et al., 2016; Shickel et al., 2018). Many studies report that EHRs are valuable knowledge sources for clinical decision support systems (Shickel et al., 2018; R. T. Sutton et al., 2020).

Clinical decision support systems can aid in various tasks, such as patient phenotyping (Sharafoddini, Dubin, and J. Lee, 2017) and disease subtyping (L. Li et al., 2015).We focus on the most crucial tasks for the differential diagnosis process.

2.1.1. Clinical Pathway Prediction

A clinical pathway is a set of therapy and treatment activities required to achieve a specific treatment objective. A clinical pathway often involves several multidisciplinary treatment activities from admission to discharge and is founded on evidence-based medical observations (Z. Huang, Dong, et al., 2014; Kinsman et al.,

2010).

Clinical Pathway Prediction is the task of assessing the patient's current situation, providing recommendations on diagnostic or treatment steps, and predicting the likelihood of clinical outcomes (Aspland, Gartner, and Harper, 2021; van Aken, Papaioannou, et al., 2021). Therefore, the CDSS needs an in-depth understanding of the patient's situation and access to medical knowledge such as clinical guidelines, medical literature, or clinical archives to derive best practices into recommendations and predictions (Aspland, Gartner, and Harper, 2021; Z. Huang, Dong, et al., 2014; Z. Huang, Z. Ge, et al., 2018; R. Liu et al., 2014).

2.1.2. Cohort Modeling

is the task of identifying patients that meet selection criteria to fit into a specified cohort, e.g., patients with type 2 diabetes (Cui, Bozorgi, et al., 2012; L. Li et al., 2015). Selecting these cohorts is essential for clinical research. Moreover, a group of similar patients is also valuable in assessing a patient's trajectory and phenotype and deciding clinical pathways (Sharafoddini, Dubin, and J. Lee, 2017). Recent studies explore the application of cohort modeling methods to discover comorbidity clusters in autism spectrum disorders (Doshi-Velez, Y. Ge, and Kohane, 2014), personalized clinical decision-making (Bellazzi, Ferrazzi, and Sacchi, 2011; Landi et al., 2020), and improvements in recruiting patients for clinical trials (Cui, Bozorgi, et al., 2012; Miotto, L. Li, et al., 2016; Sarmiento and Dernoncourt, 2016).

2.1.3. Types of Clinical Decision Support Systems

R. T. Sutton et al., 2020 performed a meta-analysis of the benefits, risks, and success strategies of CDSS ranging from 1980 until 2018. Thereby they categorize the approaches by scope, for example, Patient Safety (Eslami et al., 2012; Mahoney et al., 2007; McEvoy et al., 2017), Clinical Management (McMullin et al., 2004 Sep-Oct; Salem et al., 2018), Diagnostic Support (Cui, Bozorgi, et al., 2012; De Fauw et al., 2018; Ronicke et al., 2019) or Patient Decision Support (Jungmann et al., 2019). Besides this, they categorize the literature into rule-based (Cui, Bozorgi, et al., 2012) or trained models (Arandjelović, 2015; Bakator and Radosav, 2018; Miotto, F. Wang, et al., 2018). We extend this classification system by shedding

light on the modality of the used data. Clinical decision support systems rely on structured features, medical imaging, or text data. We will describe each data modality in the following sections.

2.1.3.1. Medical Imaging-based CDSS

Medical Imaging-based CDSS approaches analyze unstructured medical images in the medical modalities of radiology, nuclear medicine, and ultrasound (Born et al., 2021; Gamble et al., 2021; Jadhav et al., 2020; Oakden-Rayner et al., 2017). Most recent approaches to this task rely on deep-learning-based CNN (LeCun and Bengio, 1995) and GAN (Goodfellow et al., 2014) models that emerged from the field of computer vision. As a result, a substantial part of these approaches focuses on disease-specific diagnosis support due to the task's visual pattern-recognition nature. For example, such models recognize breast cancer markers (Gamble et al., 2021) or assess cardiac problems (Esteva et al., 2021; Georgiou et al., 2011; Giardino et al., 2017; N. Zhang et al., 2019)

2.1.3.2. Structured Features-based CDSS

Structured Features-based CDSS examine structured and semi-structured data, such as ICD-10 codes (Glicksberg, Miotto, et al., 2018; Landi et al., 2020), laboratory results (L. Li et al., 2015; W. T. Li et al., 2020; J. Sun et al., 2012), or time series (Cui, Bozorgi, et al., 2012). The proposed methods use the structured data as discrete input for classification or clustering methods (Cui, Bozorgi, et al., 2012; L. Li et al., 2015) or convert them into a latent representation beforehand (Landi et al., 2020; Miotto, L. Li, et al., 2016). In contrast to medical imaging-based approaches, structured feature-based approaches apply to various tasks such as selecting patients into relevant cohorts for research (Landi et al., 2020; L. Li et al., 2015), precision medicine or modeling the patient trajectory (Miotto, L. Li, et al., 2016).

2.1.3.3. Text-based CDSS

Text-based CDSS use the crucial context information recorded in free-text clinical narratives to make predictions. These systems aim to capture patients' clinical

observations recorded as narrative text, including radiology reports, operative notes, and discharge summaries, which comprise a significant portion of a patient's EHR (Demner-Fushman, Chapman, and McDonald, 2009). For example, Rajkomar et al., 2018 create neural embedding vectors for clinical free-text as additional input for predictive models. van Aken, Papaioannou, et al., 2021 follow a similar approach using BioBERT (J. Lee, Yoon, et al., 2019) as a foundation. Ronicke et al., 2019 propose a system that uses physicians' free-text input to assist in diagnosing rare diseases. Other approaches use clinical narratives for clinical information retrieval (Koopman, Cripwell, and Zuccon, 2017), cohort selection (Sarmiento and Dernoncourt, 2016; Sharafoddini, Dubin, and J. Lee, 2017; Zhu et al., 2014) or, medical coding assistance (Bell, Jalali, and Mensah, 2013; Catling, Spithourakis, and Riedel, 2018; Shi, 2017). A crucial clinical task that can benefit from the analysis of EHRs is differential diagnosis. This task involves critically exploring the patient's history and physical examination and carefully reviewing the data obtained in laboratories and diagnostic image settings (Altkorn, 2020; Croskerry, 2009).

2.1.3.4. Summary

An ideal clinical decision support system should examine all modalities to make its predictions. This challenge is often not addressed and remains primarily unsolved due to the complexity of each data modality. Much related work focuses on a specific disease (De Fauw et al., 2018; Goldenberg, Nir, and Salcudean, 2019; D. Jiang et al., 2020). We aim for a generally applicable system that uses machine learning models.

2.2. Information Extraction

Information extraction methods capture knowledge from free-form text resources. Commonly, they are used to extend or construct knowledge bases[1] and databases, or as a preparatory step in more complex text understanding applications. For example, in: *'An MRI revealed a C5-6 disc herniation with cord compression...'* such a system should extract the following two relations (Q. Wei et al., 2020):

[1] `https://tac.nist.gov/tracks/index.html`

1. TestRevealsProblem(MRI, C5-6 disc herniation)
2. TestRevealsProblem(MRI, cord compression)

Often, those methods require multiple pre-processing steps such as tokenization, part-of-speech tagging, dependency parsing, and recognition and linking of entities, which we briefly discuss in the following paragraphs.

2.2.1. Tokenization

Tokenization is the task of segmenting a sequence of characters into a semantically useful sequence of characters called tokens (C. D. Manning, Raghavan, and Schütze, 2008). C. D. Manning, Raghavan, and Schütze, 2008 elaborate that "a token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing." Tokens can be whole words adhering to the Penn Treebank standard (Marcus, Santorini, and Marcinkiewicz, 1993) or subword-oriented (Devlin et al., 2019; Jurafsky and James, 2021; Sennrich, Haddow, and Birch, 2016; Y. Wu, Schuster, et al., 2016). Tokenizers need to be efficient since tokenization is usually the most basic pre-processing step applied to all NLP tasks. A word-oriented tokenizer has to handle punctuation, special characters, and the ambiguity of words correctly. On the contrary, trained tokenizers such as Byte-Pair-Encoding (Sennrich, Haddow, and Birch, 2016), SentencePiece (Kudo and Richardson, 2018), or WordPiece (Y. Wu, Schuster, et al., 2016) split a text into units best suited for their downstream application.

2.2.2. Part-Of-Speech Tagging

Part-Of-Speech (POS) Tagging is the process of assigning a part-of-speech tag to each word in a text (Jurafsky and James, 2021). The task of a POS tagger is to disambiguate words using their context so that the correct POS tag is applied. For example, "*book*" can be a verb (*book* that flight) or a noun (hand me that *book*) (Jurafsky and James, 2021). The tagger, for example, has to decide which of the 45 tags in the English-specific Penn Treebank tagset (Marcus, Santorini, and Marcinkiewicz, 1993) is correct.

2.2.3. Dependency Parsing

Dependency Parsing aims to determine grammatical relations between words in a sentence. These relations are usually binary and consist of a head and a dependent. The parser assigns to the edge created between the words a dependency relation type such as nominal subject (NSUBJ), direct object (DOBJ), or others provided by, for example, the Universal Dependency typeset (De Marneffe et al., 2014; Nivre et al., 2016). The resulting dependency tree provides valuable information for many applications, such as clinical concept recognition, open information extraction, or question answering.

2.2.4. Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying words mentioning real-world instances of, e.g., persons, organizations, or locations. A standard approach is formulating entity recognition as a span detection problem addressed by sequence labeling (Jurafsky and James, 2021). A specific case of NER is Clinical Concept Recognition, which specializes in identifying words that mention clinical concepts in a text (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017; Si et al., 2019; Y. Wu, M. Jiang, et al., 2018). For example, in '*The patient reports a history of cancer in her family.*' an NER model recognizes the term cancer, but this term is ambiguous and may refer to multiple concepts such as 'breast cancer' or 'colon cancer' (Jurafsky and James, 2021; Schumacher, Mulyar, and Dredze, 2020). Therefore, the found concept mention needs to be disambiguated and associated with unique identifiers in standardized ontologies, such as UMLS (Bodenreider, 2004). This task is known as *named entity linking* and, in this specialized case, called *clinical concept linking* (Aronson and Lang, 2010 May-Jun; Fu et al., 2020).

2.2.5. Classical Information Extraction

Classical Information and Relation Extraction Methods apply extraction rules relying on the aforementioned intermediate structures to transform the unstructured free-form text into predefined schemas (Kilias, Löser, and Andritsos, 2015; Krishnamurthy et al., 2009). These approaches are easy to debug, permit the user a high level of direct control over the extraction process and can outperform

machine-learning based models (Chiticariu, Y. Li, and Reiss, 2013). This method requires multiple NLP system passes over the target text. Moreover, it assumes all relations' schema and extraction rules are known a priori. As a consequence, many information extraction systems are domain-dependent. Worse, the Zipfian nature of natural language (C. Manning and Schutze, 1999) causes this approach to lose significant amounts of recall. These two problems negatively affect downstream applications (Etzioni et al., 2011; Pink, Nothman, and Curran, 2014).

2.2.6. The Open Information Extraction Paradigm

The Open Information Extraction (OIE) Paradigm aims to solve these problems by capturing all relations from heterogeneous texts in a single pass without pre-specifying which schemata or special extraction rules (Banko et al., 2007; Etzioni et al., 2011). Subsequently, OIE strives for the three following goals: (1) domain independence, (2) unsupervised extraction, and (3) scalability to large amounts of text (Del Corro and Gemulla, 2013; Niklaus, Cetto, et al., 2018). OIE modifies the relation extraction task to achieve these goals. Instead of matching a predefined schema, OIE systems search for relational predicates and extract *n* associated arguments (Akbik and Löser, 2012; Del Corro and Gemulla, 2013; M. Mausam, 2016). Some task formulations include extracting additional semantic features, such as the subject of the relation, clausal modifiers, or attributions (Gashteovski, 2020; M. Mausam, 2016; Stanovsky, J. Michael, et al., 2018). For example, for the sentence: '*Since then, she had a "massive headache", which did not resolve with Tylenol.*' the model of Stanovsky, J. Michael, et al., 2018 yields the following relations:

1. had(Since then [argm-tmp], she [subject], a "massive headache" which did not resolve with Tylenol)
2. resolve(a "massive headache" [subject], which [R-subject], not [argm-neg], with Tylenol)

Researchers explore the capabilities of the open information extraction paradigm and the resulting intermediate structure in numerous downstream tasks, including question-answering (Fader, Zettlemoyer, and Etzioni, 2013; Khot, Sabharwal, and Clark, 2017; Z. Yan et al., 2018), information retrieval (Boden et al., 2011;

Kadry and Dietz, 2017), slot filling (Angeli, Premkumar, and C. D. Manning, 2015; Soderland et al., 2013; D. Yu, L. Huang, and H. Ji, 2017), knowledge base population (Lin et al., 2020; Wolfe, Dredze, and Van Durme, 2017), clinical concept linking (X. Wang et al., 2018), biomedical literature search (Q. Li et al., 2018; X. Wang et al., 2018), and biomedical knowledge graph construction (Finlayson, LePendu, and Shah, 2014).

2.2.6.1. Rule-based Open Information Extraction

Rule-based OIE Methods rely on extraction rules handcrafted by experts. These rules rely on semantic and syntactic annotations provided by NLP systems such as POS taggers or Dependency parsers. For example, ReVerb (Fader, Soderland, and Etzioni, 2011) uses POS-based regular expressions to describe its extraction rules. Other approaches (Akbik and Bross, 2009; Akbik and Löser, 2012; Angeli, Premkumar, and C. D. Manning, 2015; Christensen, Soderland, Etzioni, et al., 2011; Etzioni et al., 2011; Fader, Soderland, and Etzioni, 2011; Gashteovski, Gemulla, and del Corro, 2017; M. Mausam, 2016; Niklaus, Bermeitinger, et al., 2016; Niklaus, Cetto, et al., 2018; Pal and Mausam, 2016; Saha, 2018; Stanovsky, Dagan, and Mausam, 2015) such as ClausIE (Del Corro and Gemulla, 2013), utilize grammatical knowledge to formulate extraction patterns based on dependency parses. While delivering high precision and domain independence, covering all relevant extraction rules to achieve high recall is labor-intensive. Systems like OllIE (Mausam et al., 2012) or WOEparse (F. Wu and Weld, 2010) use bootstrapping methods to learn extraction patterns, starting with handcrafted seed rules, in a semi-supervised manner (Del Corro, 2016; Gashteovski, 2020; Niklaus, Cetto, et al., 2018). Since these approaches rely on an NLP pipeline, they suffer from recall loss similar to NEL systems, as stated in the analysis of Pink, Nothman, and Curran, 2014.

2.2.6.2. Deep-learning-based Open Information Extraction

Deep-learning-based Methods vastly increase the capability of OIE systems. The recent successes in transfer learning using word embeddings (Mikolov, K. Chen, et al., 2013; Pennington, Socher, and C. Manning, 2014) and transformer-based language representations, such as BERT (Devlin et al., 2019), are valuable foun-

dations for deep-learning-based OIE systems (Kolluru, Adlakha, et al., 2020; Kolluru, Aggarwal, et al., 2020). Recent literature explores three variations of the formulation of OIE as a deep-learning task (Kolluru, Aggarwal, et al., 2020). Sequence-labeling-based approaches such as RnnOIE (Stanovsky, J. Michael, et al., 2018) and SenseOIE (Roy et al., 2019) assign a label to every token in a sentence based on the previous decisions and input. Cui, F. Wei, and Zhou, 2018, and M. Sun et al., 2019 formulate OIE as a text generation problem and use sequence-to-sequence learning to solve it. Zhan and H. Zhao, 2020 and T. Jiang, T. Zhao, et al., 2020 formulate OIE as a span selection problem:

1. Their model predicts the position of the predicate span begin and end.
2. They generate possible spans for arguments of the relations.
3. They filter the arguments by constraints and assign each remaining argument to a relation.

As for many NLP tasks, deep-learning-based approaches are becoming the de facto standard for state-of-the-art OIE models. The generalizability and advances in transfer learning allow such models to exploit knowledge captured in pretraining tasks, for example, language modeling. We will discuss the fundamentals and capabilities of such neural text representations in more detail in Section 2.3.

2.3. Neural Text Representation

Representing natural language text for machine learning models is a non-trivial task. Text representations need to be computationally efficient yet cover all relevant aspects of language, such as semantics, syntax, global context, and local context. In this section, we first revisit the distributional hypothesis, which forms the theoretical foundation of this research branch. Next, we describe attempts at discrete and latent text representations. Subsequently, we discuss the impact of contextualized text representations such as Elmo (M. E. Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Brown et al., 2020). The biomedical domain provides unique challenges for text representations, which we discuss next. We also revisit text representation approaches specific to the biomedical domain. After presenting benchmarks and probing tasks, we investigate alternative pretraining

tasks to learn powerful text representations covering diverse textual modalities. Next, we discuss methods to combine differing text representations. Finally, we show recent applications of neural text representations and summarize the relevant related work.

2.3.1. Distributional Hypothesis

Capturing the meaning of words in an efficient representation is crucial for any language processing method. Current statistical and machine learning approaches rely on the idea that its co-occurrences in a text collection can capture the meaning of a word (Harris, 1954). Firth, 1957; Firth, 1961 formulated the distributional hypothesis in 1957 as: "a word is characterized by the company it keeps." Many researchers picked this concept up, reformulated it slightly, and found empirical proof (Rubenstein and Goodenough, 1965; Schütze and Pedersen, 1995) that there is a correlation between distributional similarity and meaning similarity (Sahlgren, 2008). This correlation allows utilizing the distributional similarity to estimate the meaning of words (Sahlgren, 2008).

Sahlgren, 2008 embeds the empirical distributional hypothesis into the theoretical framework of linguistic structuralism. Therefore, he introduces syntagmatic and paradigmatic relations between words that de Saussure et al., 1983 initially formulated:

Syntagmatic Relations concern positioning and relate entities that co-occur in a text. Syntagmatic relations are combinatorial relations, meaning that words that enter into such relations can be combined. A syntagm is such an ordered combination of linguistic entities. For example, written words are syntagms of letters, sentences are syntagms of words, and paragraphs are syntagms of sentences (Sahlgren, 2008).

Paradigmatic Relations concern substitution and relate entities that do not co-occur in the text. Paradigmatic relations are substitutional relations, which means that linguistic entities have a paradigmatic relation when the choice of one excludes the choice of another. A paradigm is thus a set of such substitutable linguistic entities (Sahlgren, 2008).

Sahlgren, 2008 concludes that distributional approaches to meaning acquisition rely on syntagmatic and paradigmatic relations between words. Finally, he formulates the refined distributional hypothesis:

> "A distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words." (Sahlgren, 2008)

We will briefly examine models that build upon the refined distributional hypothesis in the following. Furthermore, we will discuss discrete statistical models, the first approaches to applying the distributional hypothesis. With the reemergence of neural network-based models, these statistical models laid the groundwork for neural natural language models, an essential building block for many NLP systems.

2.3.2. Statistical Text Representations

Researchers based the first attempts to capture the meaning of words on co-occurrence statistics. A simple model of this type is the bag of words (Harris, 1954; Sahlgren, 2008). It holds word occurrence counts inside a context window and represents present words with a non-negative integer. A word-document matrix can describe this representation where each row represents a word and each column a document in which it occurs. Bag of words is a simple method to capture syntagmatic relationships between words (Sahlgren, 2008) using discrete features. The vector space resulting from the word-document matrix can determine the similarity between words and documents using a distance function, e.g., the inner product (Salton, Wong, and C.-S. Yang, 1975; Salton, C.-S. Yang, and C. T. Yu, 1975). When the occurrence counts for two vectors are identical, the angle will be zero, producing a maximum similarity measure (Salton, Wong, and C.-S. Yang, 1975; Salton, C.-S. Yang, and C. T. Yu, 1975).

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2.1}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \tag{2.2}$$

$$\text{score}(D, Q) = \text{tf}(t, d) \cdot \text{IDF}(q_i) \tag{2.3}$$

Salton, Wong, and C.-S. Yang, 1975; Salton, C.-S. Yang, and C. T. Yu, 1975, improve this simple model by adding weighting terms based on term frequency (Equation 2.1) and the inverse document frequency (Equation 2.2). The TF-IDF weighting term enriched the captured document-local context information with corpus-wide observations, helping to find discriminative terms (Equation 2.3).

A notable downside of this approach is the generated vector space model's sparseness and inability to capture words' meaning adequately.

Statistical language models represent language by the conditional probability of the next word given all the previous ones in a context window (Equation 2.4). Discrete statistical language models often aim to capture the meaning of single words in the vocabulary (Bengio et al., 2003).

2.3.3. Distributed Text Representations

Vector spaces resulting from discrete statistical language models are often sparse. Worse, when the number of values each discrete variable can take is extensive, most observed objects are almost maximally far from each other in the hamming distance (Bengio et al., 2003). Therefore, Bengio et al., 2003 propose a neural language model that uses neural networks to learn a smooth distribution function of word sequences. They associate a learned continuous real-valued vector with each word in a lookup table to represent word similarity. In contrast to "bag of words" models, this feature vector is often much smaller and independent of the vocabulary size. Bengio et al., 2003 use a neural network model that maximizes the parameter theta for the log-likelihood of co-occurring words (w) in the training

data to learn these feature vectors (Equation 2.4).

$$L = \frac{1}{T} \sum_{t} \log f\left(w_t, w_{t-1}, \cdots, w_{t-n+1}; \theta\right) + R(\theta) \tag{2.4}$$

The neural language modeling paradigm allows a high level of generalization (Bengio et al., 2003; Bojanowski et al., 2017; Brown et al., 2020; Devlin et al., 2019; M. Peters et al., 2017), capturing meaningful syntactic and semantic regularities (Mikolov, Yih, and Zweig, 2013; van Aken, B. Winter, et al., 2019). Building on this model, Mikolov, K. Chen, et al., 2013 propose the continuous bag of words (CBOW) and skip-gram training regimes for their word2vec model:

CBOW , unlike the standard bag-of-words model, uses a continuous distributed representation of the context of a word. The CBOW model trains to predict a word given a context window. The correct words for similar contexts probably differ in an extensive training data set, emphasizing the paradigmatic relationship between words.

The Skip-gram Regime trains to predict a word based on another word in the same sentence and emphasizes syntagmatic relations between words. The model uses each current word as an input to a log-linear classifier with a continuous projection layer and predicts words within a specific range before and after the current word.

Le and Mikolov, 2014 extend the word2vec framework to represent sentences and paragraphs with their paragraph vectors model. Pennington, Socher, and C. Manning, 2014 present a similar approach for capturing meaningful semantic and syntactic information about words with GLOVE. Bojanowski et al., 2017 extend word2vecs' skip-gram model with character n-grams to enable the model to capture the internal structure of words. Therefore, they represent each word as a bag of character n-grams and calculate the final word vector by summing over all bag elements. Introducing sub-word units into the neural language model paradigm was crucial for handling rare words and out-of-vocabulary situations. Moreover, the captured morphological properties of words are beneficial when handling morphological-rich languages (Bojanowski et al., 2017) or biomedical

texts (J. Lee, Yoon, et al., 2019; Y. Zhang et al., 2019). Another approach to deal with out-of-vocabulary words is to learn a subword representation in a data-driven way. For example, WordPiece (Y. Wu, Schuster, et al., 2016) and Byte Pair Encoding (Sennrich, Haddow, and Birch, 2016) learn subword representations that segment words using a set of character sequences of variable size. These representations learn to minimize the vocabulary size while maintaining a high language-model likelihood of the training data. This method balances characters' flexibility and the efficiency of words (Y. Wu, Schuster, et al., 2016).

2.3.4. Contextualized Distributed Text Representations

A significant drawback of the approaches mentioned earlier to text representation is their lookup table-like nature. In those approaches, every word has a single vector representing it without considering the context of the current text. For example, in the following sentences, the word "system" is associated with the same word vector regardless of the context it appears in:

> "The lymphatic system, [...] is an organ system."[1]

> "A clinical decision support system can help medical professionals."

Consequently, this single-word vector must capture all the different meanings of the word "system," leading to non-optimal performance. However, high-quality text representation should ideally model both complex characteristics of words (e.g., syntax and semantics) and how these characteristics vary across different syntagmatic and paradigmatic contexts (M. E. Peters et al., 2018). M. E. Peters et al., 2018 shifted the paradigm of obtaining word vectors from learned lookup tables towards assigning each token a representation that is a function of the entire input sentence. This approach is comparable to the Skip-Thought sentence representation proposed by Kiros et al., 2015. Many researchers (Brown et al., 2020; Devlin et al., 2019; Gu et al., 2021; Radford, J. Wu, Child, et al., 2019; Raffel et al., 2020) follow this idea with slight variations to create high-performing base models using unsupervised pretraining objectives. The resulting models deliver text understanding capabilities that transfer-learning methods, such as model fine-tuning (Devlin et al., 2019), can transform into classification performance.

[1] Wikipedia, Lympathic system: `https://en.wikipedia.org/wiki/Lymphatic_system`

The transformer architecture has recently become widespread and successfully applied for large pre-trained language models (Vaswani et al., 2017). Contrary to the approach of M. E. Peters et al., 2018; M. Peters et al., 2017 and Kiros et al., 2015, who use (bi-directional) LSTMs (Z. Huang, W. Xu, and K. Yu, 2015)) with forget gates (Gers, J. A. Schmidhuber, and Cummins, 2000). Despite the widespread success of large language models in generalizing various text-related tasks (A. Wang, Pruksachatkun, et al., 2019; A. Wang, Singh, et al., 2018), researchers still observe a domain dependence (Gu et al., 2021; Peng, S. Yan, and Lu, 2019). Consequently, researchers propose domain-specialized pre-trained language models, e.g., biomedical texts (Gu et al., 2021; J. Lee, Yoon, et al., 2019; Peng, S. Yan, and Lu, 2019). These domain-specialized models often perform better in handling biomedical entities or dealing with the characteristics of clinical narratives. As reported by Devlin et al., 2019; Mikolov, K. Chen, et al., 2013; M. E. Peters et al., 2018; M. Peters et al., 2017; Sahlgren, 2008, the information captured about words by a text representation highly depends on the pretraining task. Mikolov, K. Chen, et al., 2013 Skipgram-based model focuses more on paradigmatic relationships, while their CBOW training goal emphasizes learning syntagmatic relations. M. E. Peters et al., 2018; M. Peters et al., 2017 note the shortcomings of context in text representations and propose the bi-directional language model pretraining goal. Here, a model learns to take a larger context window into account and condition the vectors of a token on past and future words. Devlin et al., 2019 propose masked language modeling as an alternative. In this setting, a model needs to predict a masked word given future and past tokens. Additionally, they note that models often lack to learn multiple sentences spanning relationships. To combat this problem, they suggest next-sentence prediction as an additional simultaneously learned pretraining task.

2.3.5. Specialized Text Representations

While the approaches mentioned earlier focus on variations of the language modeling pretraining goals, other researchers explore task-specific text representations. These alternatives capture different textual modalities since they need a different focus to capture the information required to solve their pretraining task.

2.3.5.1. Topic Modeling

Topic Modeling is the task of discovering the discussed topic in documents or other granularities. Commonly, this task was accomplished using probabilistic models, such as Latent Dirichlet Allocation (LDA) (Blei, 2012; Blei, A. Y. Ng, and Jordan, 2003). This model spans a latent vector space in which topically similar portions of text form topical clusters. Building on the idea of latent word and document vectors, J. Liu et al., 2016 represent documents with vectors of closely related domain keyphrases. Bhatia, Lau, and Baldwin, 2016; Dieng et al., 2017 propose neural networks for topic-modeling-inspired tasks to obtain neural topic embeddings. Inherently, these models focus on the aboutness of a textual unit and need to focus on more extended contexts.

2.3.5.2. Entity Modeling

Entity Modeling is the process of representing entities or concepts in a text. The resulting representation helps disambiguate entity or concept mentions, e.g., medical conditions in clinical narratives (Choi, Chiu, and Sontag, 2016; Schumacher and Dredze, 2019; Schumacher, Mulyar, and Dredze, 2020). Kiela, C. Wang, and Cho, 2018 use latent entity representation for record linkage. Y. Ji et al., 2017 enrich the neural language model with an entity prediction training goal. This task formulation requires the neural network to condition entity vectors on the current textual context. It, therefore, emphasizes the model to capture additional paradigmatic relations in the resulting entity vector space. Similarly, Ling et al., 2020 propose to use the masked language modeling training goal to learn long-range paradigmatic relations. M. Chen et al., 2019 propose EntEval, a specialized benchmark for entity focus text-representations.

2.3.6. Holistically Capturing Textual Modalities

It is ongoing research to benchmark neural text representations for their general performance on natural language understanding tasks (M. Chen et al., 2019; A. Wang, Pruksachatkun, et al., 2019; A. Wang, Singh, et al., 2018), captured linguistic properties (Conneau, Kruszewski, et al., 2018; Ethayarajh, 2019; Köhn, 2015), and domain specificity/generality (Gu et al., 2021; Peng, S. Yan, and Lu,

2019). To excel in all these benchmarks is the goal of holistic text representations, which would allow a general base model for all text-related machine-learning tasks. Multiple branches of research investigate methods to achieve this goal. We categorize these branches into three categories: *implicit learning*, *explicit learning*, and *explicit combination*.

Implicit Learning is the idea that a model, often a language model, will learn a holistic text representation solely by observing enough text examples. Many large language models such as BERT (Devlin et al., 2019), ELMo (M. E. Peters et al., 2018), T5 (Raffel et al., 2020), GPT2 (Radford, Narasimhan, et al., 2018), and GPT3 (Brown et al., 2020) fall into this category. While implicit learning approaches have advanced the boundaries for many language understanding tasks, work in the explicit learning category reports some shortcomings.

Explicit Learning formulates specific pretraining tasks intending to enable the model to capture the information required to solve various tasks, which might require focusing on grammatical structure, local entity contexts, or document-wide context. Often this is done in multi-task learning (Sanh et al., 2022; J. Wei et al., 2022) settings. A downside of this approach is that a reasonable amount of training data for the specific pretraining tasks needs to be available, which is contrary to the self-supervised language modeling task or explicit combination approaches.

Explicit Combination of text representations aims to apply a combining function on two or more learned text representations that preserve each base representation's properties. This line of research is inspired by multi-modal representations such as image-text embeddings (Balaneshin-kordan and Kotov, 2018; Jain et al., 2021; Mroueh, Marcheret, and Goel, 2015). Building on the hypothesis that differing pretraining goals and data requires the model to capture differing text modalities, researchers aim to combine complementary text embedding to achieve a holistic representation (Coates and Bollegala, 2018; Kiela, C. Wang, and Cho, 2018; Muromägi, Sirts, and Laur, 2017; Rettig, Audiffren, and Cudré-Mauroux, 2019; L. Wu et al., 2018; Yin and Schütze, 2015).

2.4. Discussion

In this chapter, we have discussed clinical decision support systems. We have investigated typical tasks, challenges, and approaches of text-based CDSS. To be successful in supporting medical professionals, a text-based CDSS needs to capture a broad understanding of the case at hand and might exploit clinical archives as a knowledge base.

Traditional information extraction methods like the Open Information Extraction paradigm, often rely on discrete syntactic and semantic processing combined with expert-written extraction rules. These approaches might deliver valuable knowledge representations given the large scale of typical clinical archives.

Another approach to dealing with the challenges of clinical text is neural text representation. Unlike traditional information extraction methods, these latent text representations enable decision-making models to access a broader scope regarding information density and context information. Furthermore, we argue that neural text representations can be a valuable building block in holistic multi-modal neural patient representations that enable clinical decision support as an end-to-end task.

In the scope of this thesis, we follow the traditional approaches to information extraction and clinical decision support. We evaluate the suitability of the discrete information extraction methods in the differential diagnosis process (Figure 1.1). Moreover, we design neural text representations to capture topical information that language modeling-based approaches often miss. Since a holistic understanding of a medical case is crucial for CDSS, we investigate which combination of specialized neural text representations is most beneficial. We understand these information extraction and representation methods as enablers for a holistic end-to-end differential diagnosis support system.

Chapter 3

Analysing Open Information Extraction

Deriving clinical situation awareness from EHRs requires adaptable and scalable approaches to generate actionable insights from clinical notes and other written medical reports. The Open Information Extraction paradigm allows extracting relational information between entities, concepts, and events without a predefined schema.

In this chapter[1], we approach RQ1: "Is the Open Information Extraction Paradigm Suitable for Clinical Text Understanding?" We investigate this question in a two-fold way. First, we implement an In-Database Open Information Extraction system aiming for fast execution times and integration with additional structured knowledge sources. Therefore, we implement an OIE system in the Exasol Main Memory Database System[2] (Section 3.1). We investigate the runtime performance of this

[1] This chapter was published in the following articles:
R. Schneider, C. Guder, T. Kilias, A. Löser, J. Graupmann, and O. Kozachuk (2016). 'Interactive Relation Extraction in Main Memory Database Systems'. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Vol. 26. Systems Demonstrations, pp. 103–106. (Visited on 01/03/2017)
R. Schneider, T. Oberhauser, T. Klatt, F. A. Gers, and A. Löser (2017a). 'Analysing Errors of Open Information Extraction Systems'. In: *Building Linguistically Generalizable NLP Systems*. Copenhagen, Denmark
R. Schneider, T. Oberhauser, T. Klatt, F. A. Gers, and A. Löser (2017b). 'RelVis: Benchmarking OpenIE Systems.' In: *International Semantic Web Conference (Posters, Demos & Industry Tracks)*

[2] `https://www.exasol.com/`

system. Moreover, we demonstrate its capabilities to combine existing structured knowledge sources with extracted relational tuples.

Secondly, we analyze the quality of Open Information Extraction Systems in an integrated benchmark (Section 3.2). We create a conclusive benchmark using four[1] publicly available datasets (Section 3.2). We base our evaluation use cases on news analytics and supply chain risk management due to a lack of annotated datasets at the time of writing. (Section 3.2.2) We perform a quantitative evaluation using automated measurements and a manual qualitative analysis. (Section 3.2.3) We report scores for commonly known error classes and introduce "Out of Scope" as an additional error class (Section 3.2.4). In Section 3.2.5, we introduce our benchmarking toolkit and give insights on the system design and a walkthrough of our evaluation procedure. Finally, conclude this chapter in section 3.3 and review the posed research question.

3.1. Open Information Extraction in Main-Memory Database Systems

We present INDREX-MM, a main-memory database system for interactively executing two inter-woven tasks: declarative relation extraction (Krishnamurthy et al., 2009) from text and their and downstream analysis with SQL. INDREX-MM simplifies these tasks for the user with powerful SQL extensions, executing open information extraction and integrating relation candidates with domain-specific data. We demonstrate these functions on 800k documents from Reuters RCV1 with more than a billion linguistic annotations and report execution times in the order of seconds.

3.1.1. Introduction

Relation Extraction (RE) is the task of extracting semantic relations between two or more entities from text as defined in Section 2.2.5. The resulting relations are often loaded into a relational database system for further processing.

[1] Which were available at the time of publication. After publishing this chapter, the research community published new annotated datasets (T. Jiang, Zeng, et al., 2021; Kuebler, Tong, and M. Jiang, 2021; Stanovsky and Dagan, 2016).

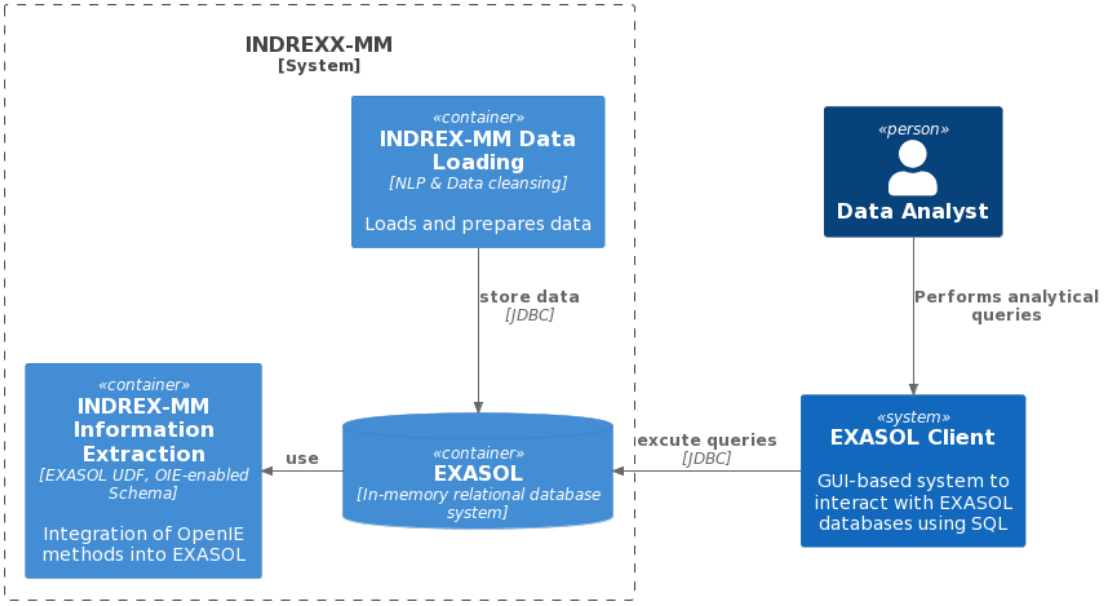


Figure 3.1.: C4 Container diagram[1] illustrating the architecture of INDREX-MM.

Use Case and Task Description: While browsing news, a supply chain analyst researches product recalls of suppliers of a car rental company. She desires to complement an existing table productrecall(supplier, product) with relations extracted from news text. Currently, the user performs these tasks with two separate systems: a system for extracting a relation productrecall(supplier, product), such as described by Krishnamurthy et al., 2009, and a relational database management system (RDBMS) for joining, grouping, aggregating and ordering. In a typical workflow, the user ships existing tables from the RDBMS to bootstrap text extraction systems and returns extracted relations to the RDBMS for analytical queries. This costly workflow is iterated until an analytical query reveals the desired insights. Moreover, the user must learn to manage both systems.

System Description. Ideally, users could execute analytical and relation extraction tasks in a single database system and leverage built-in query optimizations. Another crucial requirement is interactive query execution, particularly for extracting rare relation types with high recall and precision. We demonstrate INDREX-MM, a Main-Memory Relational Database System (MM-RDBMS) that permits this functionality as a fast backend for interactive relation extraction applications, such as

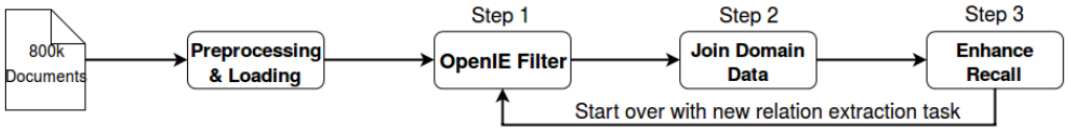


Figure 3.2.: Relation Extraction process using Open Information Extraction in INDREX-MM.

in (T. Michael and Akbik, 2015), clinical decision support systems like (Schmidt et al., 2021) or on the command line.

INDREX-MM provides broad and powerful SQL-based query operators for relation extraction. These include query predicates for detecting span proximity, predicates for testing overlapping spans or span containment, scalar functions for returning the context of a span, or user-defined table-generating functions for consolidating spans. Further, the system supports executing regular expressions and built-in operators from the RDBMS, such as joins, unions, or aggregation functions. These additional operators permit the user basic operations for looking up words in sentences describing entities or other potential relation arguments. The system also supports the user learning about potential open relation candidates where these words appear or about distributions of potential synonymous relation names. Finally, we support the user in investigating new relations. We follow the data structure design concepts of Kilias, Löser, and Andritsos, 2015, who discuss details in-depth and report extensive performance evaluations.

INDREX-MM relies on EXASOL[1], a parallel main-memory and column-oriented database. It permits integration via standard interfaces, such as JDBC, or business intelligence tools, like Tableau. The high-level system architecture is shown in Figure 3.1.

3.1.2. System Initialization

We demonstrate how INDREX-MM supports the user in three elementary steps during the declarative relation extraction process, for which figure 3.1.2 gives a high-level overview. Each of these steps 'filters out' irrelevant sentences and only keeps sentences containing relations of the type *productrecall(supplier, product)*.

[1] `https://www.exasol.com/`

Batch Loading. First, the analyst loads base annotations in a flat, sparse, and cache affine data structure. Text mining workloads rarely require full scans of all table data but do often require full scans of a small subset of the columns. Our base table layout from (Kilias, Löser, and Andritsos, 2015) supports such workflows. This schema partitions data per (document, span); we denote a span with its beginning and ending characters. Many operations on text are 'local' on a single document. Hence, our partition scheme permits an MM-RDBMS to ship data for a single document 'close' to the CPU and in orders of magnitude faster cache structures. We provide additional attributes denoting annotation types for each span, such as tokenization, sentence recognition, part-of-speech tagging, named entity recognition, user-defined types, dependency tagging, or noun- and verb-phrase chunking. We add attributes for referencing spans to containment relations in the same document. For example, a span for a sentence may contain additional spans denoting organizations. Such a flat and sparse table layout pre-joins data already at data loading time and avoids most joins at query execution time. Because of the columnar table layout in an MM-RDBMS, NULL values in attributes do not harm query execution time.

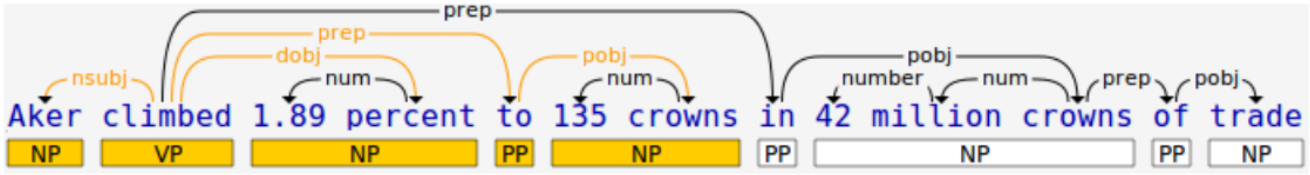
3.1.3. Filtering Relation Candidates with Open Information Extraction.

From a database perspective, we understand Open Information Extraction (OIE) (See Section 2.2) as selective filters connecting arguments in sentences. Recent work in clause-based OIE (Del Corro and Gemulla, 2013) shows effective filters for n-ary relations. INDREX-MM supports OIE as a black box or as customizable and debuggable database views: One approach is executing OIE outside an MM-RDBMS as a black box, loading results into an OIE table, and reference spans to the annotation table. We noticed that such black boxes are difficult to debug and break with the programming paradigm of the database, and if the code does not match the corpus requirements of the user, she must wait for an update of the OIE system. On the contrary, we provide the user in INDREX-MM a set of' ready-to-use OIE filters in SQL as views and user defined functions, as shown in Figure 3.3. The user can add SQL predicates from additional OIE approaches, as proposed by Angeli, Premkumar, and C. D. Manning, 2015, and can debug directly on her text corpus while the MM-RDBMS handles optimizing the execution.

```
1  select NSUBJ_VALUE, VERB_VALUE, DOBJ_VALUE, PREP_TOKEN ...
2  FROM chunks_with_id c_nsubj, verb_chunks_with_id c_verb, ...
3  WHERE
4      <char_contains("c_nsubj.span1", "dep.d_nsubj")>
5      AND <char_contains("c_dobj.span1", "dep.d_dobj")>
6      AND <char_contains("c_pobj.span1", "dep.d_pobj")>
7      AND <char_same_region("t_prep.span1", "dep.d_prep")>
8      AND <char_same_region("t_verb.span1", "dep.d_verb")>
9      AND <char_contains("c_verb.span1", "t_verb.span1")>
10     AND <char_contains("s.span1", "t_verb.span1")>;
```

NSUBJ_VALUE	VERB_V...	DOBJ_VALUE	PREP_...	POBJ_VALUE
Aker	climbed	1.89 percent	to	135 crowns
Draper	beat	's Karol Kucera	in	straight sets
The Port Authority	approved	a series	including	a runway extension
Foreigners	bought	more U.S. securities	in	Q3

(a) Extraction query and result example.



(b) Dependency parse and phrase chunks used in extractor query.

Figure 3.3.: Query example of an Open Information Extraction pattern.

3.1.4. Joining OIE Relations with Domain Data

After the first step, relations connect two or more relation arguments. However, we need to filter out irrelevant relations and keep relations of our desired type productrecall. For example, we keep relations connecting a company with predicates, such as 'recalls' and 'withdraws,' and discard relations with 'sold' or 'has refused.' For executing this task and similar to universal schemas (Riedel et al., 2013), we join arguments of OIE relations with in-house domain-specific relations representing the same semantic type, such as a table describing product recalls of a company's suppliers. As a result, our universal schema represents relations, mainly candidate patterns of our desired relation type and a few patterns for other types (see Figure 3.4a). The fast execution performance of INDREX-MM permits the user to manually filter out these irrelevant patterns. For example, she aggregates, groups, and counts patterns with standard SQL, orders patterns by frequency, and marks unsuitable patterns (see Figure 3.4). For spotting additional semantic patterns, we provide synonyms from Wordnet (Fellbaum, 1998). INDREX-MM also supports loading existing lexical patterns from the literature in a table, such as Hearst patterns (Hearst, 1992) or patterns from ConceptNet (Speer, Chin, and Havasi, 2017). The user can execute a join and utilize these patterns as additional filters for OIE candidates (see also Akbik and Löser, 2012).

Selectional Restrictions and Enhancing Recall. For further enhancing recall, the user keeps lexical patterns for predicates from the last step but applies various selectional restrictions to arguments. INDREX-MM supports selectional restrictions to one or many argument types. For example, the user may keep the company name of relations from the second step but relax the second argument. As a result, she may spot new relations of *productrecall(supplier, product)*, in particular relations between previously known companies and previously unknown products.

```
1  SELECT DISTINCT oie_pattern, nsubj_value, verb_value,
2      dobj_value, prep_token, pobj_value, verb_lemma, sentence_value
3  FROM
4      OPENIE_UNION_TABLE AS oie,
5      KNOWN_AUTOMOBILE_RECALLS AS known_recall
6  WHERE
7      oie.nsubj_value = known_recall.supplier
8      AND oie.dobj_value LIKE '%' || known_recall.product  || '%';
9
10
```

OIE_PATTERN	NSUBJ_VALUE	VERB_VALUE	DOBJ_VALUE
OpenIES4	GM	recalls	1,400 1997 Corvettes
OpenIES4	Nissan	recalls	20,500 '95 Quest minivans
OpenIES4	GM	will build	no 1998 Skylark coupes
OpenIES4	Mazda	recalls	1,560 Lancia cars

(a) Joining the Union OIE table with in-house data regarding known product recalls of the company's suppliers.

```
1  SELECT count(*), oie_pattern, verb_lemma FROM (
2      SELECT DISTINCT oie_pattern, nsubj_value, verb_value,
3          dobj_value, prep_token, pobj_value, verb_lemma, sentence_value
4      FROM
5          OPENIE_UNION_TABLE AS oie,
6          KNOWN_AUTOMOBILE_RECALLS AS known_recall
7      WHERE
8          oie.nsubj_value = known_recall.supplier
9          AND oie.dobj_value LIKE '%' || known_recall.product  || '%') as sub
10 GROUP BY oie_pattern, verb_lemma ORDER BY count(*) DESC;
```

COUNT(*)	OIE_PATTERN	VERB_LEMMA
31	OpenIES4	recall
9	OpenIES4	sell
7	OpenIES6	sell
7	OpenIES6	make

(b) Relation candidates grouped, counted, and ordered by pattern and verb. The most frequent combination is OIE-pattern four and the verb "recall."

Figure 3.4.: Use of in-house data to spot patterns of product recall mentions in the OIE schema.

3.1.5. Discussion

INDREX-MM analyzes a Billion Annotations in Seconds. We measure the relation extraction process from above in INDREX-MM on Reuters RCV1 with 800k documents and 1.2 billion annotations. For each of the four steps mentioned above, we report the execution time and how selective each filtering step prunes sentences in Table 3.1. To evaluate the accuracy, we asked two independent students to randomly draw a sample of 100 sentences after each step and count the number of correct relations for our desired type *(RL-100)*. Additionally to our product recall *(PR)* example, we report results on where a supply chain analyst wants to spot alliances *(AL)* and acquisitions *(AC)* related to a company's suppliers. We accordingly repeat the same analysis steps for these cases.

Table 3.1 shows our measurements and example sentences. One-time batch loading *(BL)* takes roughly 180 minutes, because the MM-RDBMS executes compressions and builds index structures before we can run queries. In a streaming scenario, the MM-RDBMS uses delta indexing techniques and permits hitting queries while new data is inserted. INDREX-MM exploits data locality and leverages multi-core shared memory architectures. Declarative relation extraction systems, such as SystemT (Krishnamurthy et al., 2009) or GATE4[1], need to conduct expensive data shipping between different NLP components and databases. Such data shipping is a major performance bottleneck. Contrary, INDREX-MM avoids data shipping, instead brings functionality to data, and even leverages multiple built-in optimizations of main memory RDBMSs, such as massive parallel execution with multi-cores, compression techniques, and column-based table layouts, cache affine data structures, single instruction multiple data (SIMD) or result materializations.

[1] `https://gate4.com/`

Step	Time	Relations	RL100	Example
BL	180m	15.785.155	0	-
1 OIE	9,9s	13.695.006	10	All OIE patterns(Mitsubishi, raised its production plan, October)
2 PR	49ms	134	31	Product recall(GM, recalls, 1,400 1997 Corvettes)
3 PR	619ms	921	91	Product recall(Tensor, recalls, halogen bulbs)
2 AL	16,64s	662	35	Alliance(LUKoil, signed, a $2-billion deal, with SOCAR)
3 AL	2,505s	3.265	91	Alliance(Xillix, signed, an agreement, with Olympus)
2 AC	5,643s	112	41	Acquisition(Quaker, reviews, Snapple)
3 AC	7,031s	1654	73	Acquisition(Quaker, acquired, Snapple, for, $1.8 billion)

Table 3.1.: Performance for each step. After phase *BL*, we loaded 15.7 Mio sentences and estimated one relation per sentence. In step 1, we extract OIE relations from sentences using the seven basic patterns from ClausIE, resulting in slightly fewer OIE relations than sentences. For phases 2 and 3, we show results for the relations *productrecall(supplier, product) (PR)*, *alliance(company, company) (AL)* , and *acquisition(company, company) (AC)*. We count correct relations on a randomly taken sample of 100 sentences *(RL100)*.

3.2. Analysing Errors of Open Information Extraction Systems

We report results on benchmarking Open Information Extraction systems using RelVis, a toolkit for benchmarking Open Information Extraction systems. Our comprehensive benchmark contains three data sets from the news domain and one data set from Wikipedia with an overall of 4522 labeled sentences and 11243 binary or n-ary OIE relations. In our analysis using these data sets, we compared the performance of four popular OIE systems: ClausIE, OpenIE 4.2, Stanford OpenIE, and PredPatt. In addition, we evaluated the impact of six common error classes on a subset of 749 n-ary tuples. Our deep analysis reveals important research directions for the next generation of OIE systems.

3.2.1. Introduction

Open Information Extraction (OIE) (See Section 2.2) system users often desire to select a suitable OIE system for their specific application domain. Making the right choice is a challenging task. Unfortunately, there is surprisingly little work on evaluating and comparing results among different OIE systems. Worse, most OIE methods utilize proprietary and unpublished data sets. In most cases, users can only rely on publications and need to download, compile, and evaluate existing systems on proprietary data sets.

Ideally, one could compare different OIE systems with a unified benchmarking suite. As a result, a user could identify "sweet spots" of each system but also weaknesses for common error classes. The benchmarking suite should feature diverse gold annotations with several thousands of annotated sentences. By exploring results and errors, the user can learn how to design the next generation of OIE systems or combine several systems into an ensemble.

First, We report the results of a quantitative analysis of four commonly used OIE systems: Stanford OpenIE (SIE) (Angeli, Premkumar, and C. D. Manning, 2015), OpenIE 4.2 (OIE)[1], ClausIE (CIE) (Del Corro and Gemulla, 2013), and PredPat (PP) (A. S. White et al., 2016). We omit INDREX-MM in this selection since its extraction patterns are similar to ClausIE, and we expect the results to be comparable. We evaluate the selected systems on 4522 sentences and 11243

[1] `https://github.com/allenai/openie-standalone`

Name	Type	Domain	Sent.	# Tuple
NYT-222	n-ary	News	222	222
WEB-500	binary	Web/News	500	461
PENN-100	binary	Mixed	100	51
OIE2016	n-ary	Wiki	3200	10359

Table 3.2.: Data sets in RelVis

n-ary gold standard tuples.

Second, We share in-depth insights on a qualitative error analysis of *749* n-ary tuples in *68* sentences from four gold standard data sets annotated by all four OIE systems.

Third, We design RelVis, an integrated benchmarking system for OIE systems consisting of three news data sets: NYT-222, WEB-500 (Mesquita, Schmidek, and Barbosa, 2013), PENN-100 (Y. Xu et al., 2013), and a large OIE benchmark from Newswire and Wikipedia (Stanovsky and Dagan, 2016).

3.2.2. Data Sets

Our evaluation process for Open Information Extraction systems should be convenient and comparable. To meet this goal, we design a unified data model that enables the user to perform quantitative comparisons and extensive analyses on widely used data sets. We used in our experiments four data sets, see Table 3.2, of which two feature only binary relations with two arguments. Data sets NYT-222 and OIE2016 also contain n-ary relations. These labeled data sets originate from Mesquita, Schmidek, and Barbosa, 2013 and Stanovsky and Dagan, 2016.

3.2.3. Measuring OIE Systems

A naive way to match a tuple to a gold standard is an **equal match**. The equal match strategy requires the boundaries of all arguments and the predicate to be equal with the gold standard annotations. The number of arguments must match as well. This approach delivers exact results for computing precision. However, it penalizes other, potentially correct, boundary definitions beyond the gold standard. Dealing with multiple OIE systems and their different annotation styles requires a

less restrictive matching strategy.

A second strategy is a **containment match**, where an argument or predicate is considered correct if it contains a gold standard annotation. Hence, spans from the gold standard must be fully contained inside the OIE systems' annotation spans. The number of arguments must still concur with the gold standard. This strategy may label over-specific tuples as correct. However, it is still penalizing binary systems on n-ary data sets.

Therefore, we introduce a **relaxed containment strategy** which removes a penalty for wrong boundaries especially for over-specific extractions. This strategy counts an extraction correctly, even when the number of arguments does not match the gold standard. For example, Stanford OIE, which only returns binary OIE tuples, performs well on *NYT-nary (b)*, an n-ary data set, and yields large parts of relatively short sentences as one argument. With the relaxed matching strategy, Stanford OIEs' binary extractions are counted correctly as long as they contain all gold standard arguments.

The approach of Mesquita, Schmidek, and Barbosa, 2013 has simplified the task by replacing all entities in the test set with the words "Europe" and "Asia." In our opinion, this decision is contrary to the definition of OpenIE given by Banko et al., 2007 which describes OIE as "*domain-independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus.*" and may hide or even cause problems in the analyzed systems.

Measurements. Our quantitative evaluation calculates precision, recall, and F_2 at the sentence level. Following Pink, Nothman, and Curran, 2014, we choose F_2 instead of F_1 because it gives the recall a larger impact. The basic intuition is that a high recall of an OIE system is critical to the performance of any downstream application that can apply additional filters.

3.2.4. Common Error Classes

Authors of OIE systems distinguish among six major error classes. Table 3.4 reports errors for the four surveyed OIE systems. In the following sections, we describe each error class in detail.

3.2.4.1. Wrong Boundaries

Banko et al., 2007 describe the "Wrong Boundaries" error as too large or too small boundaries for an argument or predicate of an OIE extraction. Errors in used intermediate structures, such as dependency parses or overestimation of boundaries, might cause this. Incorrect boundaries for relation arguments can prohibit fusing, linking, or aggregating tuples for the same predicate. Consequently, an additional system must filter out incorrect boundaries, which may cause a drastic recall loss.

A solution proposed in the literature is to 'wait' until intermediate systems, such as dependency parser, POS tagger etc., provide an improved generalization. However, this may not always be the case for niche domains, such as medical text or text in enterprise scenarios, where often no labeled corpora exist for intermediate systems.

Example. Consider the following example sentence: "*DENVER BRONCOS signed LB Kenny Jackson, DT Garrett Johnson and CB Sam Young*." An OIE system might emit the binary relation *signed(DENVER BRONCOS, LB Kenny Jackson, DT Garrett Johnson and CB Sam Young)*. At the same time, the gold standard data set expects the more granular n-ary relation tuple *signed(DENVER BRONCOS, Kenny Jackson, Garrett Johnson, Sam Young)*. Accordingly, we count the extraction as wrong and as a boundary error.

3.2.4.2. Redundant Extraction

Without a schema, OIE systems output redundant extractions for the same sentence, such as for the same subject-predicate structure. For example, in the sentence "Additionally, we included some other relevant results from the 2005 survey in Antwerp." SIE yields two times the tuple (we, included, other relevant results). These OIE systems are tuned towards high recall and leave the decision to filter out redundant tuples to a downstream application (Del Corro and Gemulla, 2013).

Example. For the sentence, "However, they had a significantly ($P < 0.01$) lower percentage bone in the carcass (Additional file 3)." we observe the follwing two

nearly identical extractions *had(they, significantly (P < 0.01) lower percentage bone)* and *had(they, P < 0.01) lower percentage bone)*.

3.2.4.3. Uninformative Extraction

Fader, Soderland, and Etzioni, 2011 define uninformative extractions as OIE results omitting critical information. This error can be caused by improper handling of relation phrases that combine verbs with nouns, such as light verb constructions. Adding syntactic and lexical constraints may solve this problem to a certain extent.

Example. We observed that some systems emit for the sentence "*At least one potential GEC partner, Matra, insists it isn't interested in Ferranti.*" uninformative relation tuples such as *is_not(it, interested in Ferranti)*. The systems fail to resolve the co-reference and choose a wrong relation predicate due to boundary detection errors.

3.2.4.4. Missing Extraction - False Negatives

Missing extractions describe relations that were not found by a particular system. According to Fader, Soderland, and Etzioni, 2011, missing extractions are often caused by argument-finding heuristics, choosing the wrong arguments, or failing to extract all possible arguments. One example is the case of coordinating conjunctions. Other sources of this error are lexical constraints filtering out a valid relation phrase. Another source is errors in dependency parsing.

Example. Given the sentence "*Following an i.t. delivery, the incision was closed with metal clips.*" a system might fail to produce the relation expected by the gold standard: *was_closed(the incision, Following an i.t. delivery, with metal clips)*.

3.2.4.5. Wrong Extraction

Stanovsky and Dagan, 2016 consider a tuple as correct as long as it shares a specified threshold of characters with a gold annotation. However, this policy may emit large parts of a sentence as one argument and pose additional computation effort to a downstream application. We focus on sentence-level correctness (Angeli,

Premkumar, and C. D. Manning, 2015; Mesquita, Schmidek, and Barbosa, 2013) and define a tuple as correct if the following conditions are met:

1. The selected matching strategy yields a match for the predicate.
2. The number of arguments aligns with the gold standard.
3. The selected matching strategy yields a match for all arguments.

This error class is critical since it is impossible to recover from an error of this class, and it emits a wrong signal, which might trigger additional errors in downstream tasks. In extreme cases, a system might emit extractions contradicting the originating sentence.

Example. Consider the following sentence: "*In 1987, he and his wife, Pamela, moved to Mollusk, Virginia, where they ran a bed and breakfast inn at Greenvale Manor.*" We observed the following extraction *moved(he, his wife)* which is factually wrong.

3.2.4.6. Out-of-Scope

We observe in Table 3.4 that the selected OIE systems yield more correct extractions as recognized by authors of gold data sets. For these additional annotations, we introduce an out-of-scope category. This label does not indicate an error but helps us distinguish errors of gold labels and additional annotations of a particular OIE system that are not present in the gold standard. Our two judges marked an annotation in the qualitative evaluation as out of scope if it is valid and provides an information gain. No other error category is applied to the extraction if marked as out of scope.

3.2.5. The RelVis Benchmarking System

We demonstrate RelVis, a web-based OIE benchmarking suite, which supports evaluating the four selected OIE systems. RelVis also permits users to benchmark additional OIE systems via standardized interfaces. Its integrated benchmark covers all data sets mentioned earlier in this chapter. The system permits in-depth analysis of six error classes using the introduced matching strategies and standard

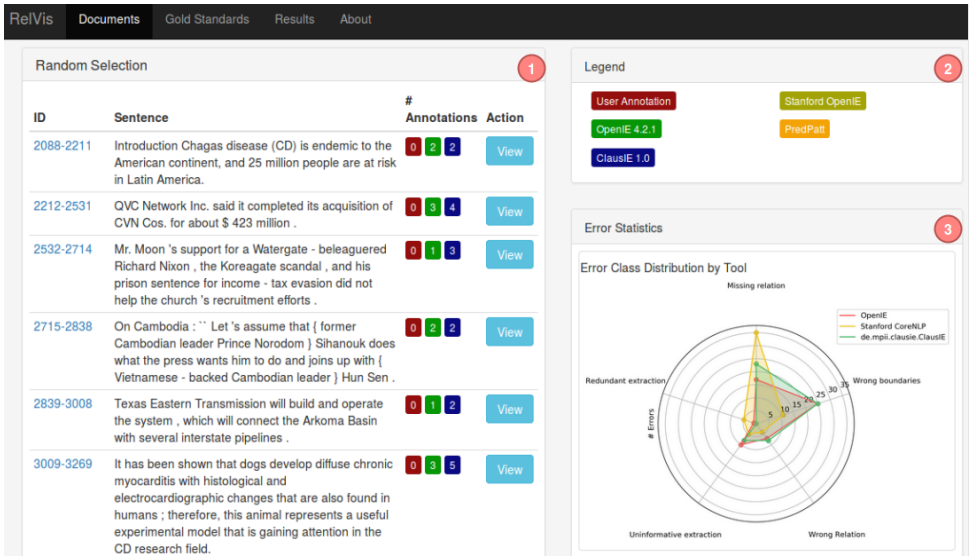


Figure 3.5.: Sentence selection view of RelVis.(1) For each sentence in the document, we show text and the number of extractions by each system. (2) Denotes various OIE systems with different colors. (3) The lower right-hand side visualizes error evaluation statistics.

quality measures, such as F-measure, precision, and recall. We used RelVis to perform all our experiments.

Startup. On system initialization, RelVis reads gold annotations and performs a quantitative evaluation. Next, the system stores extraction- and gold annotations in an RDBMS.

Dashboards for exploring annotations. Now, the user can start exploring results and understanding the behavior of each system. Figure 3.5 visualizes in a web-based dashboard sentences, precision, recall, and F scores for each OIE system and each error class.

RelVis plots error distributions as a Kiviat diagram and draws bar charts for comparing error class impacts for each OIE system. In addition, the user can export results as tables and CSV files from the database.

Managing Annotations. RelVis visualizes OIE extractions at sentence-level granularity. For each extracted relation by a system, the user can drill down into a

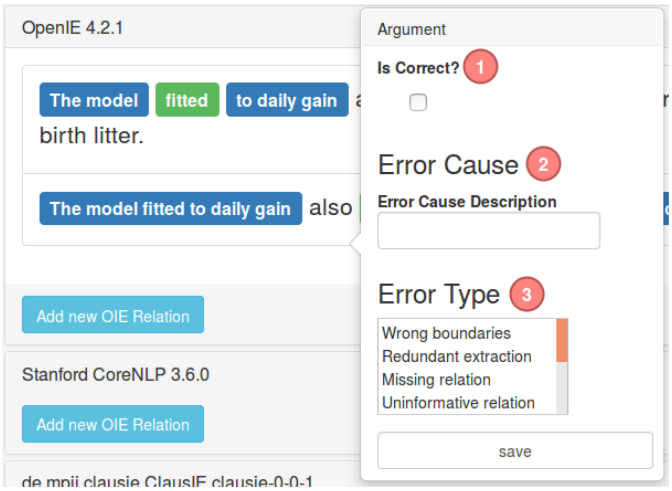


Figure 3.6.: Specifying correctness (1), error (3), and commenting on a cause (2).

single sentence and understand extraction predicates in green or arguments in blue, as shown in Figure 3.6.

Next, she can dive down into correct or incorrect annotations, add labels for error classes of incorrect annotations, or leave a comment, see also Figure 3.6. We permit the user to apply multiple error classes to each subpart of an annotation. Next, she can focus on a sentence of interest and compare extractions between different OIE systems.

The user can create them using RelVis if no gold annotations are available. Such a process is also feasible with standard annotation tools like BRAT (Stenetorp et al., 2012). However, in practice, we noted that such standard tools require many configuration steps to adapt to OIE relations. The user selects a sentence and starts with the first annotation by clicking the "Add new OIE Relation" button. Next, she marks the predicate and arguments in the sentence for her first annotation by selecting them with the cursor.

3.2.6. Experiment Results

We report precision, recall, and F_2 scores on all four data sets in a **quantitative evaluation**. Table 3.3 reports overall results for four OIE systems on all four data sets, with the limitation that only a subset of OIE2016, containing 1768 sentences,

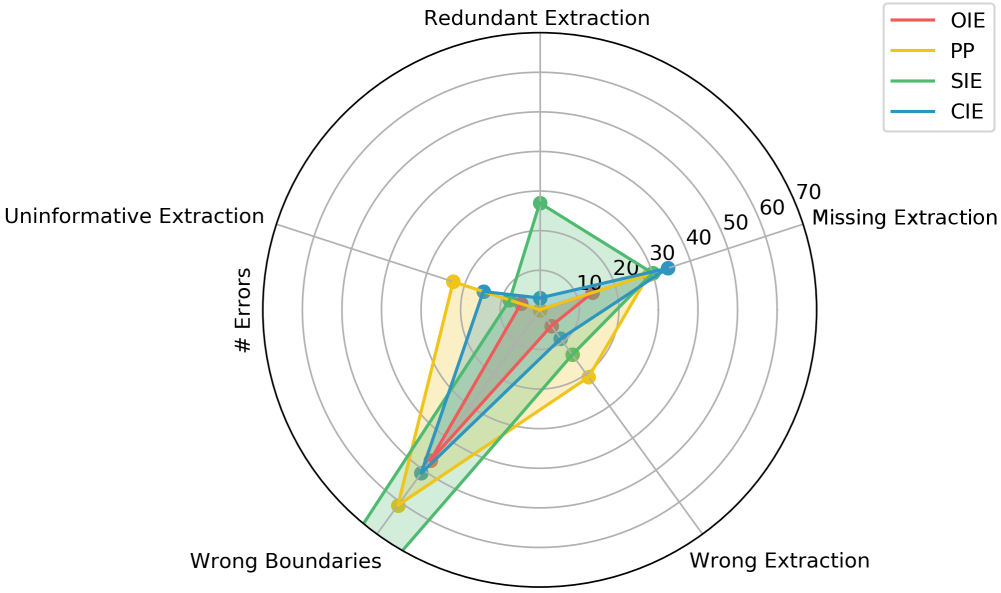


Figure 3.7.: Error occurrence of all OIE systems on 68 sentences of four data sets. Error categories described in Section 3.2.4 are plotted along five axes. The system with the smallest covered area makes the least errors. We crop the diagram at 70 occurrences for easier interpretation. SIE hits 131 times in total the Wrong Boundaries category.

was available to us. We conduct our experiments with an exact (a) and relaxed (b) containment match strategy.

We execute four OIE systems for the **qualitative evaluation** of 17 sentences of each data set. These experiments resulted in 749 predicted extractions, which we evaluate and classify into error categories by two human judges, as shown in Table 3.4. Additionally, Figure 3.7 gives an overview of the general performance of all tools over all data sets. We apply a strict containment match strategy in this evaluation. Observing that multiple errors can happen to a single extraction, we assign more than one error category in these cases.

Note, we configure system CIE to binary extraction mode for binary data sets and otherwise in n-ary mode for both experiments.

Dataset	ClausIE (%)			OpenIE 4.2 (%)			Stanford OIE (%)			PredPatt (%)		
	P	R	F_2	P	R	F_2	P	R	F_2	P	R	F_2
PENN-100 (a)	4.00	21.15	11.39	12.41	36.54	26.31	**14.85**	**57.69**	**36.58**	6.83	42.30	20.75
PENN-100 (b)	4.00	21.15	11.39	13.07	38.46	27.70	**14.85**	**57.69**	**36.59**	7.76	48.08	23.58
WEB-500 (a)	**16.33**	**46.70**	**34.03**	12.83	19.62	17.74	13.65	40.72	29.16	5.18	13.43	10.19
WEB-500 (b)	**16.33**	**46.70**	**34.03**	13.39	20.47	18.51	13.65	40.72	29.16	6.09	15.78	11.97
NYT-222 (a)	1.64	5.85	3.87	**2.86**	7.66	5.73	0	0	0	2.22	**13.51**	**6.71**
NYT-222 (b)	4.69	16.67	11.03	11.28	30.18	22.60	**13.37**	**73.87**	**38.77**	8.47	51.35	25.51
OIE2016 (a)	14.81	13.67	13.89	**24.85**	**18.69**	**19.67**	0.80	1.49	1.27	7.26	12.39	10.86
OIE2016 (b)	20.38	18.81	19.10	**39.58**	**29.76**	**31.31**	3.83	7.10	6.07	13.52	23.09	20.23

Table 3.3.: Quantitative Evaluation. The (b) variant are results with relaxed containment match strategy and (a) are those with the strict containment strategy.

Dataset	**NYT-222** (n-ary)				**OIE2016** (n-ary)				**PENN-100** (binary)				**WEB-500** (binary)			
# Relations	17				29				17				17			
	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE	CIE	OIE	PP	SIE
# Predicted	42	35	68	74	28	30	57	91	63	34	61	49	33	22	24	38
# Correct	2	1	**6**	0	8	**12**	6	5	4	8	10	**11**	5	4	3	**10**
# Redundant	0	0	0	**5**	0	0	0	**18**	1	0	0	**4**	**2**	0	0	0
# Uninformative	4	2	**8**	0	2	0	**6**	1	**9**	3	**9**	4	0	0	0	**3**
# Boundaries	11	17	18	**39**	11	11	21	**69**	**14**	5	9	**14**	8	**9**	**9**	**9**
# Wrong	2	1	3	**5**	1	1	**6**	3	3	1	**10**	4	1	**2**	**2**	**2**
# Out of Scope	24	17	**34**	30	7	6	**21**	13	**33**	17	31	18	**19**	8	12	14
# Missed	4	1	**5**	**5**	8	4	7	**12**	**14**	6	6	7	8	3	**11**	6

Table 3.4.: Occurrences of extraction errors found in the qualitative analysis of four OIE systems on 17 sentences drawn from four gold standard datasets. 749 predicted extractions were evaluated in total. Note: multiple errors per predicted extraction are possible, and that number of missed extractions is naturally not contained in # Predicted.

We observe no clear overall winner: Each OIE system works best on a particular data set, and no OIE system significantly outperforms on two or more data sets.

Boundary Errors. We observe wrong boundaries for at least one-third of the results in the four OIE systems. This result indicates that OIE systems often fail in generalizing to unseen word distributions. We observe that an OIE system causes boundary errors often by over- or under-specific argument spans. The systems emit wrong predicate spans in less frequent cases, while most are related to argument spans. Wrong intermediate structures can cause both argument and predicate-related errors. Another source of the problem could be the argument candidate generation, which overestimates the size of an argument span so that it envelops multiple distinct arguments. Further causes for a boundary error are different annotation styles, which appear among systems and gold standard data sets.

As one possible source for the overall bad results on the NYT-222 dataset, we pinpoint the differing styles of conjunction extraction. Consider a gold standard that expects a single extraction with multiple arguments for the sentence: "*DENVER BRONCOS signed LB Kenny Jackson, DT Garrett Johnson and CB Sam Young.*" e.g., *signed(DENVER BRONCOS; Kenny Jackson; Garrett Johnson; Sam Young)*. Systems CIE and OIE yield persons and their positions as one large argument in a binary relation: *signed(DENVER BRONCOS; LB Kenny Jackson, DT Garrett Johnson and CB Sam Young.)*. On the contrary, System PP implements another style, extracting every person of the sample sentence in its own binary relation.

SIE, a binary extraction system, performs surprisingly well on this data set with the relaxed containment match strategy and on NYT-222 (b). With a strict containment match strategy, NYT-222 (a), the system was not able to find a correct extraction because the data set does not contain binary relations. Using a relaxed containment match strategy, system SIE outperforms all other systems by extracting large, over-specific arguments. This behavior shifts additional effort for further processing toward downstream applications and shows the importance of considering boundaries in an evaluation. However, system SIE fails to perform on OIE2016, which contains more complex sentences, including numerical values and multiple gold annotations, compared to NYT-222.

Missed Extractions. Noisy text, wrong intermediate structures, and different annotation styles among gold data sets often trigger this error. We report a significant drop in recall for all systems on the WEB-500 dataset compared to PENN-100, except for CIE, see Table 3.3, even though both data sets show a similar annotation style. However, the WEB-500 data set is quite noisy and contains HTML-character encodings, unfinished sentences, or headlines with special characters. Those artifacts cause errors in intermediate structures, like dependency parses or POS tags, which causes the systems to fail. In particular, the n-ary systems OIE or PP do not seem to be robust to such noisy data.

Another source for missed relations is a mismatch between annotation styles. For example, system CIE shows a different style as the gold annotation in PENN-100, NYT-222, and WEB-500 data sets. A closer inspection reveals that CIE's verb-centric extraction behavior handles nominal or adjectival-triggered relations (Peng, Torii, et al., 2014) in a different style than the gold standard data set. Its design triggers inserting an artificial predicate (Del Corro and Gemulla, 2013) which can cause many missed annotations in our evaluation. Consider the following sentence: "*At least one potential GEC partner, Matra, insists it isn't interested in Ferranti.*" System CIE extracts the tuple: *is(one potential GEC partner; Matra)*, but the style of the gold standard expects *partner(GEC; Matra)*. We explain the increase of all scores of system CIE by the larger number of gold annotations compared to PENN-100, which does not interfere with the annotation style of system CIE.

Overall, we observe a trade-off among OIE systems between utilizing lexical constraints for filtering out uninformative tuples and creating false negatives. Our results indicate that system OIE handles this trade-off better than other systems.

Wrong Extractions. Wrong extraction errors are often complex and caused by other errors. For example, a boundary error often leads to missing essential information like a negation. Furthermore, we observe problems in the predicate candidate selection process for unary extractions, leading to wrong extractions.

Uninformative Extractions. Systems CIE and PP mostly yield uninformative extractions. These errors are often triggered in possessive relations without resolved co-references or relations with adjectival triggers. To overcome these problems, we suggest improving filtering for uninformative unary relations, supplying additional

checks for missed negations or important arguments, and integrating co-reference resolution components into next-generation OIE systems.

Redundant Extractions. exclusively occur in systems SIE and CIE[1]. In extreme cases, the OIE system SIE returns up to 140 tuples for the same sentence. Our results indicate that this error class has been mainly resolved in most systems by filtering and aggregating results from multiple similar extraction rules.

OIE Systems Are still Designed Towards Binary Tuples. The very first OIE systems had been designed to emit binary OIE tuples. Therefore, we observe that all systems achieve a better recall score on the binary data sets when the strict containment strategy is used. A larger number of possible errors in an n-ary task causes this—additionally, inconsistent extraction styles for n-ary relations in both systems and gold standards cause errors.

Out of Scope. The PENN-100 data set supplies for every sentence just one gold standard extraction. In most cases, it represents a non-verb-triggered relation. Since most systems perform well in extracting relations triggered by verbs, this leads to many out-of-scope extractions. Every surveyed OIE system yields out-of-scope extractions, particularly on the NYT-222 data set, which shows that the gold annotations in this data set do not cover the capabilities of modern OIE systems.

Evaluating the OIE2016 dataset results in the lowest number of out-of-scope extractions overall. It provides multiple gold annotations per sentence and covers a wide variety of extractions, starting with unary up to 7-ary tuples. System PP yields non-verb-triggered unary extractions more often than other systems, which is the reason for its steady high number of out-of-scope extractions.

[1] in binary extraction mode

3.3. Conclusion

In this chapter, we have approached RQ1, analyzing the suitability of Open Information Extraction systems for challenging domains such as clinical text (See Section 1.2). We presented INDREX-MM, a fast in-database open information extraction system. INDREX-MM enables executing Open Information Extraction at the scale of hundreds of thousands of documents with execution times in the order of seconds. We showed that such a system could efficiently integrate pre-existing knowledge with insights from text.

While Clinical use-cases require high scalability in runtime, it is also crucial to maintain high precision and recall. Therefore, we designed RelVis, a comprehensive benchmarking tool to assess the quality of OIE systems. RelVis was the first benchmark that combined four labeled datasets and supported the five most recent[1] OIE systems. RelVis allows performing both qualitative and quantitative analyses.

Using the RelVis benchmark, we revealed a lack of stringent annotation policies, making a comparative analysis and design of OIE systems challenging. Moreover, we observed that each tested OIE system depends on syntactic taggers that often propagate errors toward the logic for extracting OIE tuples. Generally, we find that systems make more errors when extracting n-ary relations. The benchmarked systems often extract unnormalized relation tuples that do not leverage the well-researched concept of "normal forms" in database theory (Codd, 1970).

Concludingly, we find that the surveyed OIE systems lack benchmarks on a wide variety of datasets and rely heavily on discrete linguistic features. Our target domain is clinical narratives, often containing lexical and grammatical errors and requiring vast semantic domain adaption (Leaman, Khare, and Lu, 2015; Starlinger et al., 2017). As a result, we argue that the reviewed OIE systems, already overfitting and struggling with news datasets, will likely surface additional issues for application in idiosyncratic domains such as clinical narratives.

[1] At the time of writing of (Schneider, Oberhauser, Klatt, et al., 2017a; Schneider, Oberhauser, Klatt, et al., 2017b)

Chapter 4

Neural Text Representations for Clinical Applications

Text-based Clinical Decision Support Systems need to capture local and global contexts to bridge the gap between text understanding and patient representation. The preceding chapter concluded that discrete text representations using the OIE paradigm have issues with achieving this goal (RQ1). Therefore, we survey the capabilities of distributed neural text representations in this chapter[1] by addressing RQ2: "*Can neural text representations aid text-based clinical decision support systems?*" and RQ3: "*Are text representations trained with differing pretraining goals complementary?*".

We study if universal language modeling-based and "specialized" text representations complement each other. Addressing RQ3 and RQ2, we benchmark the capabilities of combined and individual text representations in section 4.1, focusing on medical tasks. Subsequently, we demonstrate which combinations form a holistic relationship and improve benchmark results. We employ SentEval (Conneau and Kiela, 2018), a comprehensive benchmarking suite for comparing text embeddings, to conduct our study.

[1] This chapter was published in the following article:
R. Schneider, T. Oberhauser, P. Grundmann, F. A. Gers, A. Loeser, and S. Staab (May 2020). 'Is Language Modeling Enough? Evaluating Effective Embedding Combinations'. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Vol. 12. Marseille, France: European Language Resources Association, pp. 4741–4750

This chapter is structured as follows: First, we discuss RQ3 in section 4.1. We start by giving a high-level overview in section 4.1. Next, we introduce the PubmedSection dataset to address RQ2 and discuss its design in section 4.3. In section 4.4, we present our experiment methodology and analyze and discuss results from the experiments. Finally, we conclude this chapter in section 4.5.

4.1. Introduction

Universal embeddings, such as BERT (Devlin et al., 2019) or ELMo (M. E. Peters et al., 2018), are useful for a broad set of natural language processing tasks like text classification or sentiment analysis. Moreover, specialized embeddings also exist for tasks like topic modeling or named entity disambiguation. We study if we can complement these universal embeddings with specialized embeddings. We conduct an in-depth evaluation of nine well-known natural language understanding tasks with SentEval (Conneau and Kiela, 2018). Also, we extend SentEval with two additional tasks to the medical domain. We present PubMedSection, a novel topic classification dataset focussed on the biomedical domain. Our comprehensive analysis covers 11 tasks and combinations of six embeddings. We report that combined embeddings outperform state-of-the-art universal embeddings without any embedding fine-tuning. We observe that adding topic-model-based embeddings helps for most tasks and that differing pre-training tasks encode complementary features. Moreover, we present new state-of-the-art results on the MPQA and SUBJ tasks in SentEval.

Universal embeddings, such as BERT (Devlin et al., 2019) or ELMo (M. E. Peters et al., 2018), are an effective text representation (Conneau and Kiela, 2018; Nguyen et al., 2016). Often, they are trained on hundreds of millions of documents with a language modeling objective and contain millions to even billions of parameters. These pre-trained vectors lead to significant increases in performance in various downstream natural language processing tasks (Akbik, Blythe, and Vollgraf, 2018; Joulin et al., 2017; Mikolov, K. Chen, et al., 2013; M. E. Peters et al., 2018; Radford, Narasimhan, et al., 2018). Contrary to universal embeddings, specialized embeddings for tasks like entity linking (Gillick et al., 2019; Pappu et al., 2017) or paragraph classification (Arnold, Schneider, et al., 2019) exist. Often, specialized embeddings are trained with objectives and training datasets different from universal embeddings. This circumstance raises the question of whether universal embeddings capture all useful features for downstream tasks or if specialized embeddings may provide complementary features.

Example: Clinical Decision Support Systems. Medical literature databases, such as PubMed[1] or UpToDate[2], help doctors answer their questions. These systems benefit from methods that enrich texts with semantic concepts, like entity recognition, sentence classification, topic classification, or relation extraction (Berner, 2007; Demner-Fushman, Chapman, and McDonald, 2009). Medical language is highly specialized and often ambiguous in clinical documents (Leaman, Khare, and Lu, 2015). Documents, such as medical research papers, doctors' letters, or clinical notes, are heterogeneous in structure, vocabulary, or grammatical correctness (Starlinger et al., 2017). We propose complementing universal embeddings with specialized embeddings to execute common downstream tasks for clinical decision support systems. (See Section 1.2 and 2.1) Examples are paragraph classification, subjectivity classification, question type classification, sentiment analysis, and textual similarity.

Problem Definition. We hypothesize that specialized neural text representations may complement universal embeddings. Given is a set of both universal and specialized embeddings with different pre-training tasks for the English language (see Table 4.1). These embeddings encode words, entities, or topics. Using the *SentEval*[3] (Conneau and Kiela, 2018) benchmark, we study which combination of embeddings is complementary. Thus, we investigate if universal embeddings capture the same features as specialized embeddings.

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`
[2] `https://www.uptodate.com/`
[3] `https://github.com/facebookresearch/SentEval`

Name	Pre-Training Task	Domain	Publication	Class
ELMo (EL)	Language Modeling	Web	M. E. Peters et al., 2018	Universal
BERT (BE)	Language Modeling	Web	Devlin et al., 2019	Universal
FastText (FT)	Language Modeling	Web	Mikolov, Grave, et al., 2018	Universal
Pappu (PA)	Entity Linking	Wikipedia	Pappu et al., 2017	Specialized
SECTOR (Wikipedia) (SW)	Neural Topic Modeling	Wikipedia	Arnold, Schneider, et al., 2019	Specialized
SECTOR (PubMed) (SP)	Neural Topic Modeling	Medical	Schneider, Oberhauser, Grundmann, et al., 2020	Specialized

Table 4.1.: Comparison of neural text embeddings.

Probing Embeddings with SentEval. We probe single embeddings and combinations with *SentEval* in a transfer-learning setting on nine different tasks. *SentEval* focuses on news and customer reviews. The language in these domains differs vastly from the medical domain. Moreover, *SentEval* concentrates on single-sentence evaluation, which does not fully utilize the capabilities of contextualized embedding models (Arnold, Schneider, et al., 2019; Devlin et al., 2019; M. E. Peters et al., 2018).

Novel Datasets. We tackle the shortcomings of *SentEval* by integrating the *WikiSection-Diseases*[1] (Arnold, Schneider, et al., 2019) dataset into the *SentEval* framework. WikiSection also enables an in-depth evaluation of contextualized embeddings since its paragraph classification task is multi-sentence based. As the language in CDSS resources (e.g., PubMed) differs from the Wikipedia-based *WikiSection* dataset, we propose the *PubMedSection*[2] dataset. *PubMedSection* is a novel medical topic classification dataset created with a method inspired by distant supervision.

In-depth Experimental Evaluations on 11 Tasks. We study the properties of single and combined text embeddings and their performance on the nine tasks from *SentEval* and the two medical datasets, *WikiSection* and *PubMedSection*. We examine the differences between universal and specialized embeddings and effective embedding combinations.

4.2. Types of Text Embeddings

In the following, we investigate universal and specialized embeddings shown in Table 4.1 and discuss methods for combining embeddings.

4.2.1. Universal Text Embeddings

Recently, researchers have explored universal text embeddings trained on extensive Web corpora, such as the *Common Crawl*[3] (Mikolov, Grave, et al., 2018; Radford,

[1] https://github.com/sebastianarnold/WikiSection
[2] https://pubmedsection.demo.datexis.com
[3] https://commoncrawl.org/

J. Wu, Child, et al., 2019), the *billion word benchmark* (Chelba, 2010; M. E. Peters et al., 2018) and *Wikipedia* (Bojanowski et al., 2017). Universal text embeddings often perform language modeling tasks where the model is asked to predict a missing word given a small window of neighboring words (Joulin et al., 2017; Mikolov, Grave, et al., 2018; Mikolov, Sutskever, et al., 2013; Pennington, Socher, and C. Manning, 2014). Another common task is to predict the next or masked word of a sentence given previously predicted words as context (Devlin et al., 2019; M. E. Peters et al., 2018; Radford, J. Wu, Child, et al., 2019). For the encoder-decoder architecture, Kiros et al., 2015 propose an encoder network that encodes a sequence of words in such a way that the decoder can predict the previous and the next sentence given the encoder's vector representation.

Universal embeddings vary in their granularity at the sub-word, word, or sentence level. For example, Bojanowski et al., 2017 improved the model of Mikolov, Sutskever, et al., 2013 by adding sub-word information to handle ambiguous spelling or typos. This sub-word embedding takes advantage of the fact that similarly spelled words often have similar meanings.

Universal text embeddings encode the meaning of frequent words (Devlin et al., 2019; M. E. Peters et al., 2018; Radford, J. Wu, Child, et al., 2019). However, they perform worse than domain-adapted representations in specialized domains (J. Lee, Yoon, et al., 2019; Sheikhshabbafghi, Birol, and Sarkar, 2018). Furthermore, universal text embeddings might miss essential aspects of named entities. The reason is that most training methods are based on the co-occurrence of words in relatively short local contexts. This approach hinders the models from capturing more global features of texts such as genre, topic, receiver, the authors' intention, or they miss to learn the precise meaning of a word in special domains such as medicine (J. Lee, Yoon, et al., 2019; Sheikhshabbafghi, Birol, and Sarkar, 2018).

Also, computing embedding models for highly regulated domains is often hard and not feasible due to the lack of training data (Berner, 2007; Starlinger et al., 2017) or high computational costs (Schwartz et al., 2020).

4.2.2. Specialized Text Embeddings

Neural Topic Modeling. Arnold, Schneider, et al., 2019 introduce a specialized embedding using a coherent topic modeling task for pre-training. This model en-

codes both structural and topical facets of documents (see work of MacAvaney et al., 2018) and assigns each sentence in a document a dense distributed representation of occurring latent topics (Blei, 2012). For this purpose, the model consolidates the topical structure and context over the entire document. It leverages sequence information on the granularity of paragraphs and sentences using a Bidirectional LSTM architecture (Graves, 2012) with forget gates (Gers, J. A. Schmidhuber, and Cummins, 2000). In addition, this model captures long-range topical information. However, it does not focus on disambiguating single words. Therefore, we suggest complementing universal text embeddings (disambiguation task) with neural topic models (paragraph classification task).

Neural Entity Embeddings. Pappu et al., 2017 and Gillick et al., 2019 encode meanings of entities for entity candidate retrieval and entity disambiguation tasks. The model of Pappu et al., 2017 builds on ideas of Le and Mikolov, 2014 and models an entity using local token context. It generalizes over multiple documents and co-occurrences of entities in a document with a shared neural representation. This joint approach enables the model to capture knowledge regarding entities from training data (Pappu et al., 2017). This approach delivers a vector representation for each entity mention, encodes its relatedness to other entities, and considers local context. However, such entity embeddings capture facets of named entities but might fail to encode topical structure or non-entity words. Hence, we hypothesize that entity embeddings might benefit from combining topical embeddings.

Biomedical Domain Specialization. Sheikhshabbafghi, Birol, and Sarkar, 2018 show a domain-adapted version of ELMo (M. E. Peters et al., 2018). Their contextualized word representation performs better than a general-purpose variant, even with a smaller training set. However, this model cannot generalize to out-of-domain contexts. Therefore, J. Lee, Yoon, et al., 2019 propose BioBERT, which is a BERT model adapted to the biomedical domain. They initialize this model with pre-trained weights of the original BERT. This method prevents the shortcomings of Sheikhshabbafghi, Birol, and Sarkar, 2018 approach and preserves the ability to generalize to domains other than biomedical text.

4.2.3. Combining Embeddings

Multi-modal Combinations. Previous research reports that combining embeddings with differing training objectives is effective, such as combining data representations with different modalities into a single shared vector space. For example, Heinz, Bracher, and Vollgraf, 2017 integrate customer and image data in a shared vector space and show its effectiveness for recommending products. L. Wang, S. Li, et al., 2017 combine text and image embeddings in the field of computer vision. They employ a neighborhood-preserving ranking loss to learn a non-linear mapping between image and word embeddings for image captioning tasks.

Combining Neural Text Embeddings. To the best of our knowledge, we are the first to investigate effective combinations of universal with specialized text embeddings in an extensive study on 11 tasks. In contrast, most related work focuses on novel combination methods.

Kiela, C. Wang, and Cho, 2018 and Rettig, Audiffren, and Cudré-Mauroux, 2019 study methods to automatically select universal purpose word embeddings that are best suited for a particular task. Kiela, C. Wang, and Cho, 2018 use an attention mechanism to learn a task-specific mixture mapping between multiple word embeddings dynamically. In contrast, Rettig, Audiffren, and Cudré-Mauroux, 2019 report a method to compare and rank word embeddings regarding their relevance to a given domain. Muromägi, Sirts, and Laur, 2017 learn a linear mapping to combine various word embeddings trained on the same dataset with the same method but with different random initialization into an ensemble. They use the *ordinary least squares problem* and the *orthogonal Procrustes problem* in their objective function. The method of Yin and Schütze, 2015 is similar to Muromägi, Sirts, and Laur, 2017 but employs no orthogonality constraint on the objective function. Bollegala, Hayashi, and Kawarabayashi, 2018 introduce a local linear mapping method that takes local neighborhoods into account when projecting source embeddings into a combined vector space. This method has similarities to the work of L. Wang, S. Li, et al., 2017. Coates and Bollegala, 2018 present a surprisingly effective method to combine universal embeddings by averaging word vectors and padding them with zeros to compensate for dimensionality mismatches. However, our focus lies in studying effective embedding combinations for medical

documents.

4.3. PubMedSection a Medical Topic Classification Dataset

The capabilities of contextualized embeddings cannot be measured with the SentEval framework because all its natural language understanding tasks are single-sentence-based. None of the tasks in SentEval evaluates the domain independence of the tested embeddings. To measure such embeddings and their combinations, we extend SentEval with tasks requiring tracking contexts spanning multiple sentences. Detecting coherent topics in document passages is a challenging task that requires keeping track of the overall context of a paragraph or even the whole document.

4.3.1. The WikiSection Dataset

The *WikiSection* dataset (Arnold, Schneider, et al., 2019) consists of 38k comprehensively annotated Wikipedia articles $D = (S, L, H)$ with section and topic labels L and naturally contained headings H with respect to all of its sentences S. The dataset covers up to 30 topics about diseases (e.g., symptoms, treatments, diagnosis) or cities (e.g., history, politics, economy, climate). The task is to split Wikipedia articles d_w into a sequence of distinct topic sections $L = [l_1, ..., l_n]$, so that each predicted section $l_n = (S_k, t_j, h_i)$ contains a sequence of coherent sentences $S_k = s_1, ..., s_m$, and is associated to a heading h_i, and a topic label t_j that describes the common topic in these sentences.

4.3.2. Creating the PubMedSection Dataset

We introduce *PubMedSection*, a topic classification dataset based on medical research articles. This task requires detecting and classifying structural topic facets in plain text and is inspired by the WikiSection dataset The PubMedSection dataset consists of 51,500 PubMed articles section-wise annotated with topic labels. We construct PubMedSection similar to the WikiSection dataset. We focus on the disease subset of WikiSection with section-wise annotated medical topics, which we aim to transfer to PubMed articles. Our initial PubMed collection consists of 2,142,050 articles with 29,522,566 headings. Creating the dataset includes

learning a classifier for detecting articles in PubMed similar to WikiSection and assigning labels.

Learning to Classify Relevant Articles. Labeling such a large dataset is time-consuming and costly. Following this, we annotate the PubMedSection articles using distant supervision (Mintz et al., 2009; Morgan et al., 2004) with WikiSection as ground truth. For this purpose, we filter the open-access subset of PubMed[1] D_p for articles that exhibit high textual similarity to WikiSection for a successful label transfer. We model a neural network-based non-linear binary classifier for this task[2].

First, we encode all headlines of the WikiSection diseases subset $H_w = \{h_{w1}, ..., h_{wn}\}$ as well as the headlines of the PubMed articles $H_p = \{h_{p1}, ..., h_{pn}\}$ with a fastText (Mikolov, Grave, et al., 2018) embedding model. For this step, we train a domain-specific fastText model on the full corpus of the open-access subset of PubMed. Next, we use concatenated fastText encoded word vectors of each article's headlines H_w, H_p as input for our model. We choose a one-layer neural network with ReLu activation (Glorot, Bordes, and Bengio, 2011) and softmax output over more complex architectures to minimize computational complexity. We train on 3200 human-labeled examples for the headline structure similarity task. We use the Xavier weight initialization (Glorot and Bengio, 2010), and employ Adam (Kingma and Ba, 2015) with stochastic gradient descent as an optimizer and a multi-class cross-entropy loss.

Our hyperparameter search suggests an L2 regularization (A. Y. Ng, 2004) of 10^{-4}, a learning rate of 10^{-5}, a batch size of 128, and we set the training duration to 60 epochs.

Assigning WikiSection Labels to PubMedSection. After training, we sample the top 51,500 articles by their similarity score from the filtered PubMed collection. Next, we calculate the cosine similarity between every headline of each article set (D_p, D_w) to estimate the probability that a topic label for the PubMed headline could be generated from the WikiSection labels. Next, we transfer the best-fitting topic labels from the best-matching headline's section in WikiSection

[1] `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`
[2] `https://github.com/DATEXIS/pubmedsection`

Headline	Transferred Label
Abstract	disease.other
Disease name and synonyms	disease.etymology
Definition and diagnostic criteria	disease.diagnosis
Epidemiology	disease.epidemiology
Clinical description	disease.symptom
Etiology	disease.other
Diagnostic methods	disease.diagnosis
Differential diagnosis	disease.diagnosis
Genetic counseling	disease.management
Antenatal diagnosis	disease.diagnosis
Management including treatment	disease.management
Unresolved questions	disease.other
Psycho social concerns	disease.other
Conclusion	disease.other

Table 4.2.: This table shows the result of the label transfer process applied to (Connor and Thiagarajan, 2007). Headlines follow the order of the original article. The "Transferred Labels" column shows labels assigned using our structure similarity model.

to the sampled PubMed article's corresponding section (See Table 4.2). That way, every section of a Pubmed article gets labeled with 'SectionAnnotations.' These annotations have the following attributes: *begin*, *end*, *sectionHeading*, and *sectionLabel*. Begin and end refer to the covered characters' span in the overall document. sectionHeading contains the heading of the section at hand, and sectionLabel is one of the WikiSection.disease classes. For example, the section "Disease name and synonyms" in (Connor and Thiagarajan, 2007) results in the following sectionAnnotation: "{*begin: "2345", end: "2517", sectionHeading: "Disease name and synonyms", sectionLabel: "disease.etymology*"}." Finally, we split the dataset into a training subset with 50.000, a validation subset with 1000, and a test subset with 500 labeled articles.

4.4. Evaluating Embedding Combinations

We evaluate the performance of embeddings as well as their combinations. Our methodology follows the paradigm of probing tasks (van Aken, B. Winter, et

Name	Embedding A	Embedding B
FT+PA	fastText	Pappu
FT+SW	fastText	SECTOR (Wikipedia)
FT+SP	fastText	SECTOR (PubMed)
EL+FT	ELMo	fastText
EL+PA	ELMo	Pappu
EL+SW	ELMo	SECTOR (Wikipedia)
EL+SP	ELMo	SECTOR (PubMed)
SW+PA	SECTOR (Wikipedia)	Pappu
SP+PA	SECTOR (PubMed)	Pappu

Table 4.3.: Surveyed embedding combinations.

al., 2019; Weston et al., 2016): We test combined embeddings on nine natural language understanding tasks and data sets from SentEval as well as two tasks from the WikiSection and the PubMedSection dataset. For probing these eleven tasks, we train a linear classifier with single or combined embeddings as input and observe the properties of different embedding types and their combinations. As the combination method, we chose *concatenation*. Despite its simplicity, concatenating embeddings is a strong baseline (Coates and Bollegala, 2018; Kiela, C. Wang, and Cho, 2018; Rettig, Audiffren, and Cudré-Mauroux, 2019; Yin and Schütze, 2015). Other combination methods are subject to our future research.

4.4.1. Surveyed Text Embeddings and Combinations

We select a variety of universal and specialized embeddings, as shown in Table 4.1 for our experiments. Our evaluation setting is sentence-based. Some of the surveyed embeddings are word vector-oriented. Therefore, we follow Arora, Liang, and Ma, 2017 and Perone, Silveira, and Paula, 2018 and average word vectors in a sentence for each of those word embeddings to obtain a sentence embedding vector. The embeddings employed are:

Random (RND) As a baseline, we compute random vectors.

fastText (FT) (Mikolov, Grave, et al., 2018) is word vector-oriented and trained on a language modeling task with word and sub-word tokens.

ELMo (EL) (M. E. Peters et al., 2018) train a bi-directional language modeling task with two stacked LSTMs that use a Character-CNN to capture sub-word information.

BERT (BE) (Devlin et al., 2019) builds on the transformer architecture (Vaswani et al., 2017) and masked language modeling task pre-training.

Entity embedding (PA) proposed by Pappu et al., 2017 is training an embedding for a named entity disambiguation task with a knowledge base as the target, like Wikidata[1] or UMLS[2].

SECTOR Wikipedia (SW) Arnold, Schneider, et al., 2019 propose a contextual topic embedding trained with section headings from Wikipedia articles. They show that the latent topic information contained in their SECTOR embedding can be utilized to segment documents and classify these segments into up to 30 topics.

SECTOR PubMed (SP) Same as above but trained on our novel PubMedSection dataset.

Embedding Combinations. We choose the combinations of embeddings presented in Table 4.3 for our experiments. We assume the most compelling improvements are obtained when combining specialized with universal text embeddings. We verify this assumption by evaluating if combining the two universal embeddings ELMo and fastText is as effective as combinations with specialized embeddings. Additionally, we conduct experiments with the combination of the entity embedding with both SECTOR models.

Embedding Models. We evaluate the following models as provided by their authors: *BERT Large* (BE), *ELMo Original 5.5B* (EL), *fastText crawl-300d-2M-subword* (FT), *Pappu* (PA) and *SECTOR SEC>H+emb@fullwiki* (SW). These models cover a wide variety of domains and topics. In contrast to our SECTOR PubMed (SP) model, we train exclusively on medical research articles.

[1] `https://www.wikidata.org`
[2] `https://www.nlm.nih.gov/research/umls/index.html`

4.4.2. Tasks and Parameters

We use SentEval (Conneau and Kiela, 2018) to perform an analysis of the effectiveness of each embedding combination of natural language understanding tasks. We integrate the WikiSection diseases and PubMedSection tasks into SentEval[1] to obtain comparable results for our evaluation.

We conducted our survey on the following nine plus two medical tasks: WikiSection and PubMedSection with ten-fold cross-validation.

Textual Similarity: MRPC Paraphrase detection (Dolan, Quirk, and Brockett, 2004) on the Microsoft Research Paraphrase Corpus, which consists of sentences pair extracted from *news* sources, is a binary classification task of deciding whether a sentence paraphrases another or not.

Textual Similarity: SICK-E Sentences Involving Compositional Knowledge Entailment (Marelli et al., 2014) is a 3-class natural language inference classification task based on sentences collected from Flickr *image captions* and the Microsoft Research *Video Description* Corpus.

Sentiment Analysis: MPQA Multi-Perspective Question Answering (Wiebe, T. Wilson, and Cardie, 2005) is a binary sentiment classification task on a *news dataset* from the world press.

Sentiment Analysis: SST-2 Stanford Sentiment Analysis (Socher et al., 2013) is a binary sentiment classification task on a *movie review* data set.

Sentiment Analysis: SST-5 Stanford Sentiment Analysis (Socher et al., 2013) is a fine-grained 5-class sentiment analysis task based on the same corpus as SST-2 (*movie review*).

Sentiment Analysis: CR Customer Reviews (Hu and B. Liu, 2004) is a binary sentiment analysis task based on *product reviews*.

[1] `https://github.com/DATEXIS/SentEval-k8s`

Parameter	Value
KFOLD	10
CLASSIFIER_NHID	0
CLASSIFIER_OPTIM	Adam
CLASSIFIER_BATCHSIZE	64
CLASSIFIER_TENACITY	5
CLASSIFIER_EPOCHSIZE	4
CLASSIFIER_DROPOUT	0

Table 4.4.: Parameters used in evaluation with SentEval as suggested by Conneau and Kiela, 2018.

Sentiment Analysis: MR Movie Reviews (Pang and L. Lee, 2005) is a binary sentiment analysis data set on *movie reviews*.

Classification: SUBJ Subjectivity vs. Objectivity (Pang and L. Lee, 2004) is a classification task of subjectivity and objectivity in *movie reviews*.

Classification: TREC Text Retrieval Conference Question Answering (Voorhees and Tice, 2000) 6 class question type classification. The corpus consists mostly of *newswire and newspaper* articles.

Coherent Topic Classification: WikS WikiSection diseases (Arnold, Schneider, et al., 2019) is a 27-class topic classification task sourced from the *medical subset of Wikipedia*.

Coherent Topic Classification: PubS PubMedSection is a novel 27-class topic classification task based on medical research articles from *PubMed*. We randomly sample the PubMedSection training set down to 2200 articles since evaluating the whole training set is prohibitively time-consuming.

Evaluation Parameters. We use the parameters provided by Conneau and Kiela, 2018, as shown in Table 4.4.

Model	Strong +	Minor +	Minor -	Strong -
Language Model combined with Topic Model				
EL+SW	**7**	2	1	1
EL+SP	**5**	3	2	1
FT+SW	**6**	3	0	2
FT+SP	**7**	2	2	0
Language Model combined with Entity Embedding				
EL+PA	0	**6**	4	1
FT+PA	**6**	5	0	0
Topic Model combined with Entity Embedding				
SW+PA	**5**	2	1	3
SP+PA	**6**	4	0	1
Language Model + Contextualized Language Model				
EL+FT	0	**8**	3	0

Table 4.5.: This table shows the effectiveness classification in tasks for each surveyed embedding combination. We count a model combination as "Strong+" if it advances in more than one percentage point in accuracy compared to both of its base models. Accordingly, we count a result as "Minor+" if the improvement is smaller than one percentage point. "Minor-" and "Strong-" are similarly defined for performance decreases.

4.4.3. Experiment Results

Table 4.5 overviews the results of seven single embeddings as well as nine embedding combinations on eleven evaluation tasks. Table 4.6 shows accuracy scores for single model performance and combined embedding models. Finally, Table 4.7 reveals the delta of each surveyed embedding combinations' score regarding their source embedding scores.

	Textual Similarity		Sentiment Analysis					Classification			
Model	MRPC	SICK-E	MPQA	SST-2	SST-5	CR	MR	SUBJ	TREC	WikS	PubS
RND	66.49	56.69	68.77	49.92	23.39	63.76	49.48	49.60	20.60	15.24	24.99
FT	69.86	74.35	86.69	78.69	39.68	72.00	74.68	90.22	76.00	39.11	28.15
PA	72.17	76.62	85.31	77.43	40.00	74.09	72.90	89.65	78.80	39.84	28.04
SW	67.71	67.30	84.30	65.68	34.80	80.34	77.56	*97.84*	69.60	29.82	29.27
SP	67.19	56.48	*96.85*	71.28	37.65	76.19	82.91	95.61	66.20	*46.84*	*39.43*
BE	69.16	75.75	86.91	***89.57***	*49.37*	***90.07***	*84.84*	95.83	***93.20***	44.94	31.12
EL	*73.68*	*79.54*	90.00	85.01	47.19	83.39	80.66	94.56	92.40	43.09	30.85
Language Model combined with Topic Model											
EL+SW	73.86	79.48	92.58	85.34	**49.59**	87.23	86.25	**99.17**	88.60	45.05	32.11
EL+SP	**74.61**	78.87	96.14	86.66	45.57	84.53	87.03	97.26	92.80	**50.86**	39.76
FT+SW	70.78	74.51	90.35	76.28	40.18	82.91	83.49	98.13	73.60	42.64	31.24
FT+SP	70.96	74.33	**97.34**	77.81	43.71	80.69	**87.11**	97.27	78.60	49.60	**39.85**
Language Model combined with Entity Embedding											
EL+PA	73.45	79.81	90.27	85.94	45.97	83.47	80.91	94.40	92.80	42.67	30.70
FT+PA	72.99	79.07	87.03	81.11	41.95	76.42	75.54	91.17	84.80	41.36	28.78
Topic Model combined with Entity Embedding											
SW+PA	71.65	74.79	89.92	75.40	40.68	82.89	83.60	98.43	77.60	43.18	31.16
SP+PA	72.70	77.15	97.16	75.95	44.34	81.14	87.05	97.36	85.80	50.13	39.63
Language Model combined with Contextualized Language Model											
EL+FT	73.33	**79.91**	90.19	85.78	46.65	83.76	80.79	94.46	93.00	43.24	30.97
SOTA	93.00[c]	87.80[d]	93.30[b]	96.80[c]	64.40[e]	87.45[a]	96.21[c]	95.70[f]	98.07[a]	56.70[g]	-

Table 4.6.: This table shows the accuracy score of single model approaches and the best embedding combinations for each task. We highlight the overall best score with **bold** numbers while numbers in *italic* denote the best single model results. Additionally, we gathered recent results on our surveys tasks in the SOTA row, which are reported by the following publications: (Cer et al., 2018)[a], (H. Zhao, Lu, and Poupart, 2015)[b], (Z. Yang et al., 2019)[c], (Subramanian et al., 2018)[d], (Patro et al., 2018)[e], (S. Tang and de Sa, 2018)[f] and (Arnold, Schneider, et al., 2019)[g] on section-wide topic classification. We do not take SOTA results into account when highlighting the best results since they are obtained with specialized models.

	Textual Similarity		Sentiment Analysis					Classification			
Comb. Δ	MRPC	SICK-E	MPQA	SST-2	SST-5	CR	MR	SUBJ	TREC	WikiS	PubS
Language Model combined with Topic Model											
EL+SWΔEL	0.18	-0.06	2.58	0.33	2.40	3.84	5.59	4.61	-3.80	1.96	1.26
EL+SWΔSW	6.15	12.18	8.28	19.66	14.79	6.89	8.69	1.33	19.00	15.23	2.84
EL+SPΔEL	0.93	-0.67	6.14	1.65	-1.62	1.14	6.37	2.70	0.40	7.77	8.91
EL+SPΔSP	7.42	22.39	-0.71	15.38	7.92	8.34	4.12	1.65	26.60	4.02	0.33
FT+SWΔFT	0.92	0.16	3.66	-2.41	0.50	10.91	8.81	7.91	-2.40	3.53	3.09
FT+SWΔSW	6.15	12.18	8.28	19.66	14.79	6.89	8.69	1.33	19.00	15.23	2.84
FT+SPΔFT	1.10	-0.02	10.65	-0.88	4.03	8.69	12.43	7.05	2.60	10.49	11.70
FT+SPΔSP	3.77	17.85	0.49	6.53	6.06	4.50	4.20	1.66	12.40	2.76	0.42
Language Model combined with Entity Embedding											
FT+PAΔFT	3.13	4.72	0.34	2.42	2.27	4.42	0.86	0.95	8.80	2.25	0.63
FT+PAΔPA	0.82	2.45	1.72	3.68	1.95	2.33	2.64	1.52	6.00	1.52	0.74
EL+PAΔEL	-0.23	0.27	0.27	0.93	-1.22	0.08	0.25	-0.16	0.40	-0.42	-0.15
EL+PAΔPA	1.28	3.19	4.96	8.51	5.97	9.38	8.01	4.75	14.00	2.83	2.66
Topic Model combined with Entity Embedding											
SW+PAΔSW	3.94	7.49	5.62	9.72	5.88	2.55	6.04	0.59	8.00	13.36	1.89
SW+PAΔPA	-0.52	-1.83	4.61	-2.03	0.68	8.80	10.70	8.78	-1.20	3.34	3.12
SP+PAΔSP	5.51	20.67	0.31	4.67	6.69	4.95	4.14	1.75	19.60	3.29	0.20
SP+PAΔPA	0.53	0.53	11.85	-1.48	4.34	7.05	14.15	7.71	7.00	10.29	11.59
Language Model combined with Contextualized Language Model											
EL+FTΔEL	-0.35	0.37	0.19	0.77	-0.54	0.37	0.13	-0.10	0.60	0.15	0.12
EL+FTΔFT	3.47	5.56	3.50	7.09	6.97	11.76	6.11	4.24	17.00	4.13	2.82

Table 4.7.: This table shows the delta in the accuracy score of each model combination with respect to the respective single model accuracy score. We highlight numbers in green if an embedding combination yields improved scores compared to both source embeddings.

4.4.3.1. Language Models plus Topic Models

We observe a significant increase in accuracy scores in 35 out of 44 experiments (see Table 4.5), which qualify in 25 cases for the "Strong+" category when combining a language modeling-based embedding with a topic embedding. Moreover, we report EL+SP as the overall best-performing model with a macro accuracy across all tasks of 75.83. We conclude that language modeling and topic modeling pre-training tasks capture complementary information.

ELMo Plus SECTOR Yields a Substantial Increase in Accuracy. Combining of EL and SW yields "Strong+" results (see Table 4.5) for 7 of the 11 downstream tasks. We observe only a considerable performance loss of 3.8 percentage points for the TREC task. For the three other measurements of this model, the performance increases slightly for two tasks by less than 0.33 accuracy points and decreases for the SICK-E task by 0.06 accuracy points (see also Table 4.7). EL and SP also yield strong results, with five tasks for which the source models encode complementary information. We observe only one "Strong-" loss in performance of the SST-5 task of 1.62 and report for all remaining tasks a fluctuation in performance between "Minor+" and "Minor-."

Models fastText and SECTOR Encode Complementary Features. Results for fastText plus SECTOR are nearly analog to ELMo, except that we observe an even higher performance increase on average, as shown in Table 4.7. We note that tasks MRPC, SICK-E, and SST-2 do not benefit from the features captured in SW and SP. Surprisingly, the situation for the fine-grained sentiment classification task SST-5 is different compared the results of the binary sentiment analysis task SST-2. We observe a considerable accuracy increase for the model combination EL+SW and FT+SP, a small increase for FT+SW, and a performance decrease for EL+SP.

EL+SP Outperforms EL+SW in the Medical Domain. Corresponding to the differing training domains of the SW and SP model, we can observe a more substantial increase in performance for the combination of EL and SP in both medical tasks WikiSection-Diseases and PubMedSection compared to combining EL and SW. Likewise, we observe a similar situation for the combination of fastText

(FT) with SW and SP. We explain this result by the fact that SP is trained with medical research articles and closer to the target domain than SW.

New SOTA for MPQA and SUBJ Tasks. Table 4.6 shows that embedding combinations EL+SP (96.14 acc) and FT+SP (97.34 acc) outperform the current state-of-the-art in the MPQA task (see (H. Zhao, Lu, and Poupart, 2015) 93.30 acc). Similarly, EL+SW drastically outperforms the current state-of-the-art (see (S. Tang and de Sa, 2018) 95.70 acc) in the SUBJ task with 99.17 accuracy measure. Following this result, we conclude that the differing pre-training task captures complementary features that lead to improved evaluation results.

Different Pre-Training Tasks Capture Complementary Features. We verify the complementary nature of the pre-training tasks with an additional experiment. We re-evaluate the SUBJ task with a fastText model similar to SW, exclusively trained on Wikipedia (Bojanowski et al., 2017). With this setting, we control if the objective writing style in Wikipedia is the cause of our good results. We observe only a small increase in accuracy for the Wikipedia-based model (90.98 Acc) compared to the FastText model trained on the Common Crawl (90.22 Acc). Following this result, we conclude that different pre-training tasks of FT and SW capture complementary features that lead to improved evaluation results. We explain the complementary nature of these combinations with the document-wide context that topic models encode. Topic models need to keep track of the context coherently over whole documents while respecting local topic shifts. Contrary to language modeling-based embeddings that often focus mainly on local context spanning nearby sentences.

4.4.3.2. Combinations with Entity Embeddings

Table 4.5 shows "Strong+" increases in accuracy for 17 out of 44 experiments for embedding combinations that include the surveyed entity embedding (PA).

Topic Plus Entity Embeddings Outperform. We examine the combination of the topic (SW, SP) and entity embeddings (PA) in Tables 4.5 and 4.7. Intuitively, it seems reasonable to assume that topic embeddings focus more on structure than

on the meaning of single words and, therefore, capture complementary knowledge. Our results prove this assumption, with 17 out of 22 experiments showing an increase in performance and 11 scores qualifying as "Strong+." Similar to the results when combining topic and language models, we explain the performance gains with the complementary nature of the entity disambiguation and topic modeling pre-training tasks. Additionally, we note that PA does not encode any contextual information at prediction time, while SW and SP do. Following this, it is reasonable to assume that SP+PA and SW+PA are generally beneficial combinations.

Combining fastText and Pappu is Beneficial. For 6 out of 11 tasks is our complementary constraint in Table 4.5 fulfilled, the remaining tasks have a "Minor+" accuracy increase, lower than one percentage point. We observe that FT+PA is a beneficial combination since no task has a drop in accuracy.

ELMo Already Captures Features Encoded by Pappu. On the contrary, we observe no accuracy gain over one percentage point for EL+PA. We observe six times a minor increase, four times a minor decrease and one time a strong decrease. As reported in Table 4.7 this strong decrease is accounted to the SST-5 task with a loss of 1.22 percentage point compared to the single model result of EL. Overall, we observe that this combination yields results comparable to the single model performance of EL (see Table 4.6). This result suggests that the contextualized nature of EL already captures the features encoded by PA.

4.4.3.3. Baseline and Domain Transfer

To validate our results, we survey if adding more semantically meaningful dimensions to a vector is sufficient to obtain results comparable to our experiments. Therefore, we evaluate combining a contextualized (ELMo) with a traditional language model (fastText). Next, we report the results of the single model evaluation of contextualized and traditional language models on WikiSection and PubMedSection. Finally, we survey if we can enrich a universal embedding (ELMo or fastText) with domain-specific features (SP) without losing its domain independence.

ELMo Plus fastText has no Effect. We report no result which qualifies as either "Strong+" or "Strong-" in Table 4.5 for EL+FT. We observe a slight increase in accuracy in six out of nine cases, and in three cases, minor decreases. Intuitively, it is sound to assume that contextualized embeddings (EL) should not benefit from static word embedding (FT) methods. Correspondingly, we evaluate, on the one hand, the combination EL+FT in order to investigate this intuition and, on the other hand, to obtain a baseline. Consequently, we conclude that adding semantically meaningful dimensions to text representations alone is insufficient to achieve good results comparable to the other surveyed combinations.

Classical Embeddings Perform Surprisingly well in Multi-sentence Tasks. Table 4.6 reports surprisingly good results for non-contextualized embeddings (FT and PA) in the WikiSection and PubMedSection tasks. Their best results on WikiSection (39.84 acc) and PubMedSection (28.15 acc) are quite close to the contextualized universal embeddings BE and EL (WikS: 44.94 acc, PubS: 30.85 acc). These results contradict our initial assumption that the contextualized embeddings would vastly outperform FT and PA on multi-sentence-based tasks.

Domain Specificity. We observed 18 times a "Strong+" increase in accuracy for the 33 experiments that involve SP, which is trained on PubMed abstracts (see Table 4.7 and Table 4.5). Therefore, we can confirm the observation of J. Lee, Yoon, et al., 2019 and Sheikhshabbafghi, Birol, and Sarkar, 2018 that in-domain text representations perform better on biomedical texts than universal representations. Moreover, we can show that it is possible to transfer the domain adaption into a combined embedding without experiencing catastrophic forgetting since we only observe three out of the 18 "Strong+" increases in the medical tasks (WikS, PubS). For example, as shown in Table 4.6 the combinations of EL+SP and SP+PA deliver the best results in our evaluation for the WikiSection disease task while being in the top three surveyed embedding combinations.

4.4.4. Discussion

Adding Topic Models Helps for Most Tasks. Our results suggest that adding topic models to either language models or entity embeddings is beneficial for the

overall performance of most investigated classification tasks. This observation can be explained by the topical and structural information captured in these models. Moreover, these topical models capture the coherent flow of topics across long-range dependencies while considering local topic shifts. Therefore, neurons in these models may be able to capture long-range dependencies from long documents. This information seems complementary to information from universal text embeddings or entity embeddings, with a comparably short context window.

Textual Similarity Tasks Do Not Benefit Much. We observe for textual similarity tasks only for very few scenarios a "Strong+" improvement when combining embeddings. We argue that existing universal embeddings, such as ELMo or fastText, already represent sufficient features from local features close to the target word.

Concatenation is Simple but Easily Interpretable. Our study is limited to concatenation as the operator for combining embeddings. This simple operator has a significant disadvantage in raising the dimensionality. Additionally, it does not leverage the originating correlations in combined embedding spaces. However, despite these shortcomings, this operator permits surveying for effective embedding combinations in an explainable manner.

Different Pre-Training Tasks Encode Different Features. Our study confirms that embeddings trained with different pre-training tasks can encode complementary features. Combinations of specialized and universal embeddings often result in domain-independent performance increases.

4.5. Conclusion

This chapter surveyed neural text representations for their capabilities supporting text-based clinical decision support systems. First, We addressed RQ2 by evaluating neural text representations on our newly introduced dataset "PUBMEDSECTION," and on the "WIKISECTION.DISEASES" dataset for medical topic segmentations in German and English.

Second, we investigated if text representations with differing pretraining tasks are complementary and can yield a holistic text representation when combined (RQ3). We analyzed and identified effective combinations of universal and specialized text embeddings in an extensive study on 11 tasks. We extended SentEval to the medical domain by integrating the WikiSection.diseases and the PubMedSection task. Our comprehensive analysis shows that combining universal and specialized embeddings, such as Elmo + SECTOR, yields vastly improved results in many downstream tasks. Furthermore, we showed that complementary combinations yield holistic text representations that achieve a new state-of-the-art for two tasks in SentEval.

Overall, we conclude that the SECTOR model is a robust and extensible building block representing clinical narratives that captures features complementary to language models.

Chapter 5

Deep Learning Enabled Clinical Decision Support

Clinicians often have busy schedules and have limited time to research complicated cases in medical publications and guidelines or to consult colleagues on similar cases. A large part of the information clinicians seek is available as more, e.g., guidelines and research papers, or less, e.g., clinical notes and nursing notes in EHRs structured text data. Deep learning-enabled text-based clinical decision support systems have the potential to reduce the needed effort drastically. Following this idea, we approach RQ4: "How Effective are Deep Learning Enhanced Medical Information Seeking Processes?" in this chapter[1].

Based on our previous chapters' findings, we present *Smart-MD: IR* and *Smart-MD: DDx* to address RQ4. *Smart-MD: IR* focuses on the external information-seeking process of clinicians. We design a process that enables medical professionals to search for topical queries of the form *[disease, topic]*. Founded in this process, we draft and implement a paragraph retrieval system based on the SECTOR (Arnold, Schneider, et al., 2019) model.

[1]This chapter was published in the following articles:
R. Schneider, S. Arnold, T. Oberhauser, T. Klatt, T. Steffek, and A. Löser (2018). 'Smart-MD: Neural Paragraph Retrieval of Medical Topics'. In: *Companion of the The Web Conference 2018 on The Web Conference 2018*. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 203–206
R. Schneider, M. Mayrdorfer, H. Schmidt, K. Budde, F. A. Gers, and S. Staab (2022). 'SmartMD: Deep Learning Enabled Differential Diagnosis'. In: *To Appear*, p. 20

In *Smart-MD: DDx*, we focus on the differential diagnosis process. We formalize the differential diagnosis process and operationalize its core tasks. Accordingly, we abstract this process with high-level information system operations that enable implementation using classical and deep learning-based models for information retrieval and text understanding. We design and implement the *Smart-MD: DDx* system based on this process. In a qualitative evaluation with clinicians, we validate the deep learning-enabled differential diagnosis process and reveal design challenges for text-based clinical decision support systems.

This chapter is structured as follows: First, we discuss the medical information-seeking process on external sources, such as PubMed, in Section 5.1. We introduce an exemplary scenario as a running example in Section 5.1.1. Section 5.1.2 describes employing a topic classification model such as SECTOR and a named entity recognition model such as Tasty (Arnold, Dziuba, and Löser, 2016; Arnold, Gers, et al., 2016) for medical information-seeking tasks. Section 5.1.3 discusses, based on our running example, how a paragraph retrieval system can aid medical information retrieval.

Secondly, we focus in Section 5.2 on the differential diagnosis process, which we briefly introduce in Section 5.2.1. In Section 5.2.2, we design the deep learning-enabled differential diagnosis process. Section 5.2.3 operationalizes each process step and describes which models and methods we use. Next, describe our qualitative evaluation and observation study in Section 5.2.4. We discuss our results in Section 5.2.5 and present design challenges for clinical decision support systems. Section 5.2.6 provides additional context to our work.

Finally, we conclude this chapter in Section 5.3.

5.1. Neural Paragraph Retrieval of Medical Topics

We demonstrate Smart-MD: IR, an information retrieval system for medical professionals. The system supports topical queries in the form [disease topic], such as ["lyme disease", treatments]. In contrast to document-oriented retrieval systems, Smart-MD: IR retrieves relevant paragraphs and drastically reduces a medical doctor's reading load. We recognize diseases and topical aspects with a novel paragraph retrieval method based on bidirectional LSTM (Gers, Pérez-Ortiz, et al., 2002; Hochreiter and J. Schmidhuber, 1997) neural networks. We demonstrate

Smart-MD: IR on a dataset that contains 3,469 diseases from the English language part of Wikipedia and 6,876 distinct medical aspects extracted from Wikipedia headlines.

5.1.1. Introduction

Medical doctors, particularly in emergencies, often need to make fast decisions without thoroughly studying the latest research results from journals. In particular, less experienced doctors might overlook alternative treatments or therapies and often fall back to potentially less effective standard procedures known from their academic studies. Even though most queries of doctors are of informational intent (R. W. White and Horvitz, 2014; Yoo and Mosa, 2015), standard medical search engines, like PubMed[1], still focus on filtering documents for a keyword query. Ideally, a doctor could use an effective search engine for retrieving diverse and potentially unknown results from the latest literature about symptoms, therapies, medications, treatments, or other often requested aspects during the anamnesis.

Scenario: Consider the case of a doctor searching for treatments for *Lyme disease*, an infectious disease caused by bacteria of the Borrelia type which is mainly spread by *ticks*.She will study essential articles and find the transmission of ticks from birds to humans as the main cause. While she knows from her academic studies that antibiotics such as *doxycycline* will help most patients, she might oversee that certain patients with cardiac diseases will likely suffer from this treatment and should rather be treated with *ceftriaxone*-based antibiotics. Ideally, the system would retrieve all treatments for Lyme disease and display an aggregated overview of different treatments, including some paragraphs of text explaining infrequent edge cases.

We demonstrate Smart-MD: IR, an information retrieval system that provides such a functionality for medical professionals. It takes as input diseases and a list of optional topical aspects and returns paragraphs about the given diseases in the context of the given aspects. Moreover, it recognizes and aggregates important facets in these paragraphs, such as correlating medical terms or topics, and provides the user with these facets for query refinement. Figure 5.1 shows a typical result

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`

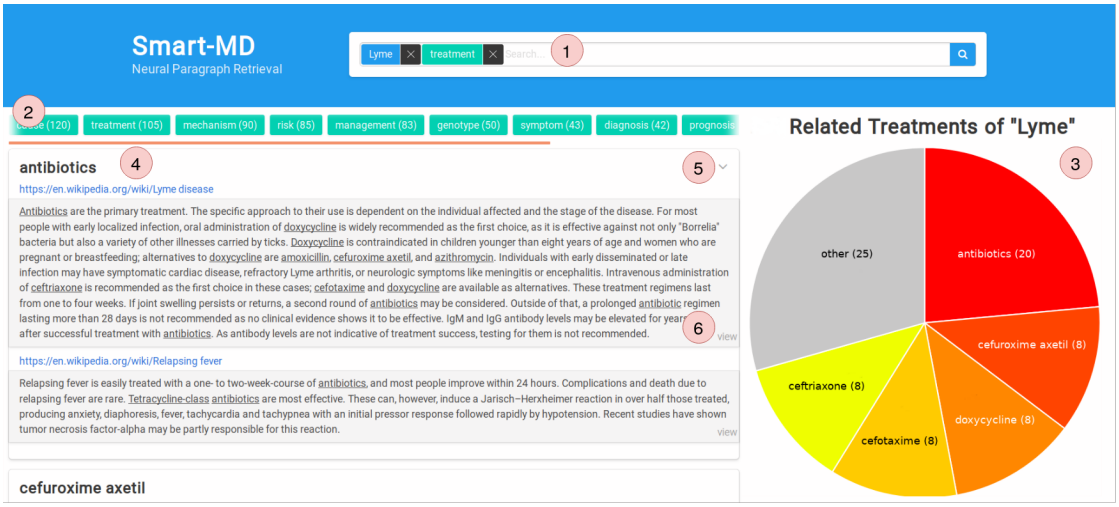


Figure 5.1.: Screenshot of the Smart-MD: IR user interface. The search bar (1) shows current query terms and offers an auto-completion based on the neural entity and topic extractors. The fact distribution chart (3) and the topic tag bar (2) offer visual navigation components to allow the user to refine the search direction. Smart-MD: IR groups search results by their topics in result cards with a generated title and short description (4). Those can be unfolded using the arrow button (5). The view button (6) opens a full-text view of the document as shown in Figure 5.2.

for the query 'lyme treatment'. Given the query (1), the system retrieves two highly relevant paragraphs about treatments from two articles on Lyme disease (4) or on Borrelia. The user is able to refine the query with topical aspects that appear in the context of these documents (2). Next, Smart-MD: IR shows a distribution of treatments (3), and the user can narrow the query to a particular novel and previously unknown treatment. Finally, the user may click on an interesting paragraph to inspect the context of the entire document. Thereby, the system highlights the topic of each relevant paragraph (6). In particular, with long documents, this fine granularity at the paragraph level permits the reader to skip many irrelevant passages.

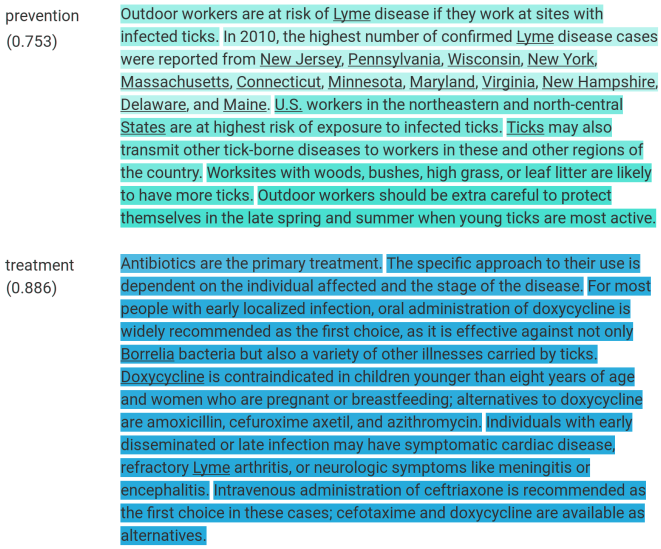


Figure 5.2.: Visualization of the neural topic classification for an example document (excerpt). Smart-MD: IR assigns coherent topic labels 'prevention' and 'treatment' to sentences. The shading of colors visualizes the confidence of the best-scored class from the prediction; numbers in brackets depict the average confidence per paragraph.

5.1.2. Paragraph Retrieval

Smart-MD: IR is built upon two neural information extractors that process the dataset at load time. The *topic extractor* assigns a distribution of topics to each sentence in the dataset. The *entity extractor* recognizes named entities in these sentences. Both models are trained end-to-end with data from the medical domain, in particular for the disease scenario. We store all extractions in an index and retrieve them at query time to return relevant paragraphs. In this section, we describe these steps briefly.

5.1.2.1. Sequential Topic Classification

The topic extractor's goal is to assign a coherent distribution of topics over all positions in a document. In contrast to traditional probabilistic topic models such as LDA (Blei, A. Y. Ng, and Jordan, 2003), which describe topic distributions at the document level, we approach capturing topics on the sentence level. One possible solution is Paragraph Vectors (Le and Mikolov, 2014), which treats all

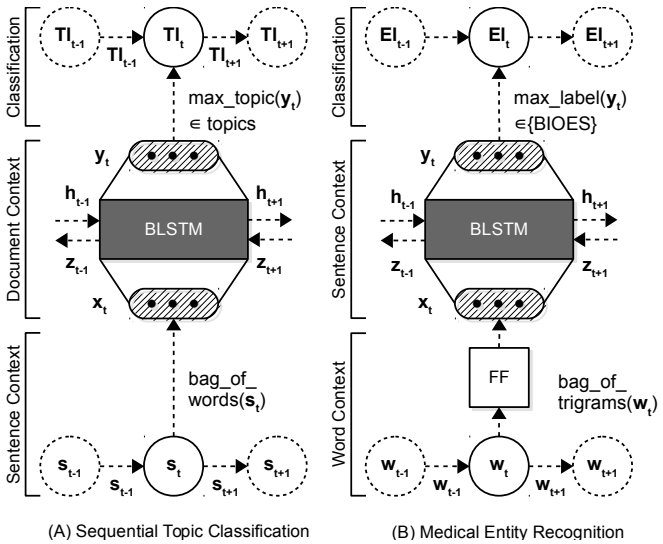


Figure 5.3.: Neural network architectures of our section classification model. It classifies a sequence of sentences $\mathbf{s}_t$ to their corresponding section labels $\mathbf{l}_t$. We employ bidirectional LSTM layers, which respect the long-range dependencies of sentences $\mathbf{s}_t$ inside a document.

paragraphs (or sentences) independently. However, to achieve a coherent sequence of topics, e.g., to spot adjacent sentences that express treatments of a disease, we need to respect the sequential order and long-range dependencies of sentences in the document. Our approach uses a Long Short-Term Memory (LSTM) network (Hochreiter and J. Schmidhuber, 1997) for classification.

Definition of Topics from Wikipedia Section Headlines. We utilize section and subsection headlines from Wikipedia documents to define possible topics. For example, we observe 6,876 distinct headlines from 3,469 Wikipedia pages on diseases[1]. Table 5.2 shows the distribution of observed topics among articles. A closer inspection reveals that this distribution is heavily skewed, e.g., the top 20 topics cover more than 90% of all paragraphs. We therefore chose 20 representative topic labels for training and assign the label 'other' to the remainder. A detailed overview of the topic distribution is shown in Table 5.1.

[1] taken from the 20170320 Wikipedia dump

topic	freq	F1	topic	freq	F1
abstract	14.35%	82.40	classification	2.29%	37.78
treatment	12.59%	70.55	genotype	2.13%	49.60
symptom	11.99%	65.59	prevention	1.69%	68.07
diagnosis	11.62%	73.43	culture	1.58%	50.24
cause	10.05%	48.98	research	1.33%	60.09
other	7.20%	23.17	animal	0.66%	50.25
mechanism	6.28%	58.05	transmission	0.63%	0.00
management	4.09%	37.78	risk	0.37%	0.00
epidemiology	4.00%	75.08	complication	0.13%	4.65
history	3.82%	66.19	screening	0.11%	0.00
prognosis	3.08%	62.80			

Table 5.1.: Distribution of covered sentences by topics in the Wikipedia dump, which was used to train the topic extractor. F1 scores are evaluated on a test set of n=32,045 sentences.

Sequential Classification using BLSTM Networks. We utilize the LSTM model with forget gates (Gers, J. A. Schmidhuber, and Cummins, 2000) and bidirectional layers (Graves, 2012) to predict for each sentence s_t a probability distribution y_t for its topic label $\mathbf{Tl}_t = \max(y_t)$. The BLSTM is configured using forward and backward layers with input nodes $\vec{g}_t$, input gates $\vec{i}_t$, forget gate $\vec{f}_t$, output gate $\vec{o}_t$ and internal state $\vec{s}_t$. We encode hidden states $\vec{h}_t$ (forward layer) and $\overleftarrow{z}_t$ (backward layer) for every time step t. We generate the output layer y_t by summing $\vec{h}_t$ and $\overleftarrow{z}_t$.

$$
\begin{aligned}
\vec{g}_t &= \phi(\vec{W}_{gx}x_t + \vec{W}_{gh}\vec{h}_{t-1} + \vec{b}_g) \\
\vec{i}_t &= \sigma(\vec{W}_{ix}x_t + \vec{W}_{ih}\vec{h}_{t-1} + \vec{b}_i) \\
\vec{f}_t &= \sigma(\vec{W}_{fx}x_t + \vec{W}_{fh}\vec{h}_{t-1} + \vec{b}_f) \\
\vec{o}_t &= \sigma(\vec{W}_{ox}x_t + \vec{W}_{oh}\vec{h}_{t-1} + \vec{b}_o) \\
\vec{s}_t &= \phi(\vec{g}_t \odot \vec{i}_t + \vec{s}_{t-1} \odot \vec{f}_t) \\
\vec{h}_t &= \vec{s}_t \odot \vec{o}_t \quad / \quad \overleftarrow{z}_t = \overleftarrow{s}_t \odot \overleftarrow{o}_t \\
y_t &= \phi(\vec{W}_{yh}\vec{h}_t + \overleftarrow{W}_{yz}\overleftarrow{z}_t + b_y)
\end{aligned}
\tag{5.1}
$$

Our network architecture is shown in Figure 5.3. We use n-hot bag of words vectors as input features, i.e., $x_t = \sum_{w \in s_t} i_w$ with indicator $i_w \in \{0,1\}^{|\mathcal{V}_w|}$ over a fixed vocabulary $\mathcal{V}_w$. We implement our BLSTM model with 300 cells, sigmoid activation, 0.5 dropout, and a softmax output layer. It is trained document-wise using stochastic gradient descent with ADAM (Kingma and Ba, 2015), L2 regularization, and cross entropy-loss using a learning rate of 10^{-3} and backpropagation-through-time (Werbos, 1990). The network classifies a complete document per iteration and is only reset in between documents. We segment the document into paragraphs by splitting it at positions where the topic label changes. The outcome of our method is visualized in Figure 5.2.

5.1.2.2. Medical Named Entity Recognition (NER)

The entity extractor's goal is to recognize medical named entities, such as diseases or medications in the documents. This task is often difficult since only sparse training data exists for this specialized task, and recall suffers (Pink, Nothman, and Curran, 2014). We utilize previous work on TASTY[1] (Arnold, Dziuba, and Löser, 2016), a generic and robust approach for high-recall named entity recognition and linking in many languages and with sparse training data. TASTY offers strong generalization over domain-specific languages, such as in biomedical text (e.g., Medline, PubMed, or Wikipedia articles), and can be trained with only a few hundred labeled sentences to achieve F1 scores in the range of 84–94% on standard datasets.

Robust recognition using character n-gram embeddings. Similar to the topic extractor, the architecture of our entity extractor utilizes a BLSTM architecture. The model's objective is to assign BIOES entity labels $\mathbf{El}_t = \max(y_t)$ to all words in a sentence (Ratinov and Roth, 2009). To achieve a robust classifier, we encode words as bag of letter-trigrams as input features, i.e., $x_t = \sum_{tri \in w_t} i_{tri}$. This allows us to train a character embedding that is able to recognize typical syllables in a word (Arnold, Gers, et al., 2016). We extract possible diseases and other medical entities and store them in the index for query completion and paragraph retrieval.

[1] Demo available at `http://demo.datexis.com/tasty/`

5.1.2.3. Query Processing and Paragraph Scoring

Smart-MD: IR executes queries of the form [disease, topic] as follows: First, the user matches ambiguous disease and topic names using autocomplete. It maps a variety of notations from Wikipedia headlines to well-defined classes. We then conduct a conjunctive boolean search and retrieve documents containing a single document's disease name and topic ID. Finally, we score the candidate paragraphs. Our scoring approach builds on the assumption that paragraphs likely contain medical entities mutually related to the paragraph's topic and the requested disease. Moreover, we would like to retrieve low-frequency events that are probably unknown to the doctor. We measure for each paragraph, proximity between the requested topic and co-occurring entities with normalized pointwise mutual information (Bouma, 2009) (nPMI):

$$\text{nPMI}(\text{entity}, \text{topic}) = \frac{\ln \frac{P(\text{entity},\text{topic})}{P(\text{entity})P(\text{topic})}}{-\ln\, P(\text{entity}, \text{topic})} \tag{5.2}$$

$P(\text{entity})$ denotes the probability that retrieved paragraphs contain the entity, $P(\text{topic})$ the probability that the topic is discussed in the retrieved paragraphs and $P(\text{entity}, \text{topic})$ denotes the probability that an entity appears in any retrieved paragraph that discusses the topic. Hence we assign to low frequency events relatively high scores.

5.1.3. Demonstration Outline

We demonstrate Smart-MD: IR in a live demonstration and with a video[1] that shows the case for our query from this chapters introduction ["lyme disease", treatments].

Initial Search Query. While she is typing the query, the system auto-completes terms against words in the index of diseases or topics. Next, the system retrieves documents, filters, scores, and displays top-ranked paragraphs. Now, she can skim the results to get an overview. The system supports her with a short description of the relevant paragraphs of the documents. All sources claiming the same fact are

[1] `https://www.youtube.com/watch?v=kcDi7qQxpBo`

no.	headline	topic	freq	H
0	Abstract	abstract	3,453	0.03
1	Diagnosis	diagnosis	2,795	0.49
2	Treatment	treatment	2,789	0.49
3	Signs and Symptoms	symptom	1,921	0.69
4	Causes	cause	1,531	0.69
		...		
14	Symptoms	symptom	339	0.32
15	Types	classification	329	0.31
16	Research	research	312	0.30
17	Society and Culture	culture	310	0.30
18	Mechanism	mechanism	224	0.24
		...		
6,873	Fungal Meningitis	other	1	0.00
6,874	Location and Symptoms	symptom	1	0.00
6,875	Molecular Basis of Disease	other	1	0.00

Table 5.2.: Frequency and entropy (H) of top-5 head and randomly selected torso and tail headings for 3,469 diseases and 6,876 distinct headlines in the English Wikipedia.

aggregated and can be unfolded by a click on the arrow icon. This representation allows her to overview and skip irrelevant content fast until she reaches interesting treatments.

Query Refinement. Smart-MD: IR ranks co-occurring entities and topics in a pie-chart or respectively in the topic bar by their frequency. If the resulting paragraphs are still too broad, she can click on topics in the topic bar to refine the query and search for rare facts. Alternatively, she can visit the entity navigation chart on the right that shows a frequency distribution of entities in paragraphs. For accessing less frequent but relevant entities that co-occur with the search query, she clicks on a pie in the chart. This excludes the more frequent entities from the visualization

and allows her to inspect the results in the 'long tail' of search results.

Inspecting the Context of a Paragraph. Finally, she can drill down into the context of interesting facts by clicking on the text that opens the corresponding document. Next, the system displays the entire document. Like hand-written notes at the margins of a textbook, Smart-MD: IR shows an assigned topic for each paragraph. She can now read these pre-labeled topics and skip topics fast until she reaches an important part. She can now drill down further or start over again.

5.2. Deep-Learning-enabled Differential Diagnosis

By its nature, the differential diagnosis process benefits from 1000s of cases an experienced doctor has seen after many years of practice. On the other hand, DDx poses a challenging situation for an inexperienced doctor at the beginning of her career. Especially, but not exclusively for their support, we propose a differential diagnosis support system called *SmartMD*. We argue that recommendations by this system can help less experienced medical professionals to improve treatment quality and optimize diagnostics usage. Primarily, it can be valuable for a doctor to obtain *complementary hypotheses* regarding possible diagnoses or subsequent clinical actions based on cases from the clinical archive (see Figure 5.4).

5.2.1. Introduction

In the last years, neural networks outperformed in diverse natural language processing (A. Wang, Pruksachatkun, et al., 2019) or computer vision tasks (Jia et al., 2021). Several authors investigate whether they can transfer these successes into clinical decision support systems. Such systems should assist medical professionals in their day-to-day work, including assessing a patient's situation, hypothesizing diagnoses, and planning diagnostics and treatments (R. T. Sutton et al., 2020). Successful examples are clinical support system approaches, such as cohort modeling (Glicksberg, Miotto, et al., 2018), outcome prediction (van Aken, Papaioannou, et al., 2021), clinical coding (Schumacher, Mulyar, and Dredze, 2020), and medical image processing (Oakden-Rayner et al., 2017).

Differential Diagnosis Process (DDx). At a high-level abstraction, DDx is a complex process of integrating symptoms, lab and vital data, and many more signals into a patient representation created in the doctor's mind. The doctor also compares this representation with similar cohorts from her experience. Thereby, she tries to spot from these' archived' cases recommendations on diagnostics that can exclude or confirm potentially severe conditions and treatments for the remaining most likely diagnoses (see Figure 5.4).

DDx: Integrate, Retrieve/Filter, Predict. Ideally, such a system solves a broad set of tasks and situations medical professionals face daily (Miotto, F. Wang, et al.,

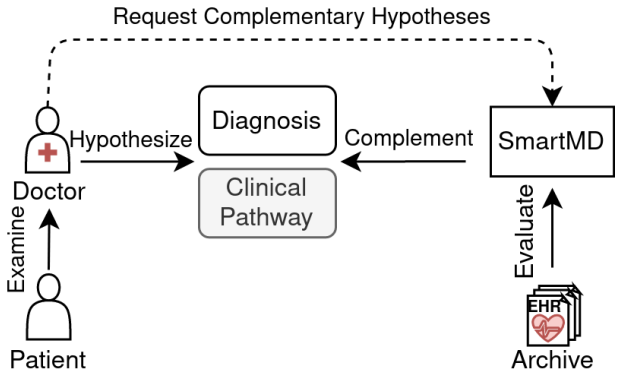


Figure 5.4.: Example of how SmartMD integrates into the clinical process. A Medical doctor examines a new patient. She further investigates the case and forms hypotheses regarding possible diagnoses and clinical pathways. To ensure that the patient gets the best possible care, she consults SmartMD to obtain complementary hypotheses.

2018; Sharafoddini, Dubin, and J. Lee, 2017; R. S. Sutton and Barto, 2016). We abstract this process with high-level information system operations. First, a system would need to *integrate* various signals from electronic health records (EHR) into a patient representation. EHRs typically include text, lab data, medical imaging, and other data describing the patient's situation. Next, the system *retrieves and filters* cases similar to the current patient from a clinical archive and groups them into cohorts. We design this process as highly interactive to incorporate the doctor's expertise. The doctor might include or exclude several aspects until retrieved cohorts *match* their assessment, the domain of the medical application, and the current patients' situation. Finally, the system needs to conduct *classification and outcome prediction tasks*, including subsequent diagnostics, diagnoses, and recommending effective treatments.

Contribution

Building such a DDx system is a complex undertaking. Together with a large university hospital in Germany and the DATEXIS research group, we designed and evaluated essential components of this system over the last years (See: (Arnold, Gers, et al., 2016; Arnold, Schneider, et al., 2019; Schneider, Arnold, et al., 2018; Schneider, Oberhauser, Grundmann, et al., 2020; van Aken, Papaioannou, et al.,

2021)). We combine individual contributions into a single system and evaluate it with medical doctors. We believe that the novelty of our integrated system and the focus on deep learning for text data could become a blueprint for the design of future DDx Systems. We discuss the following contributions in this chapter:

Abstraction of Six Core Tasks. We formalize DDx into a process comprising six tasks and three feedback loops following Altkorn, 2020; Croskerry, 2009. This framework enables us to combine the power of predictive models with the expertise and background knowledge of medical professionals as proposed by J. Sun et al., 2012.

Deep Learning & Statistical Models For DDx. We design and implement the DDx process with models for information extraction, cohort modeling, diagnosis prediction, and clinical action recommendation in an integrated system. For example, *Smart-MD: DDx* analyzes medical records using clinical concept recognition and links the extracted mentions toward the Unified Medical Language System (UMLS) (Bodenreider, 2004). Furthermore, we employ a transformer-based (Vaswani et al., 2017) negation detection method (van Aken, Trajanovska, et al., 2021) to determine if, e.g., a disease was excluded or diagnosed. Additionally, we publish the source code of SmartMD under an open-source license[1].

Evaluation with Medical Doctors. The third contribution is an in-depth qualitative evaluation of our method. The goal is to understand how our approach can be beneficial to medical professionals in their day-to-day work. We conduct the evaluation process through a multi-faceted user study. Therefore, we ask medical professionals to work with the system on randomly drawn cases from MIMIC-III (Johnson et al., 2016) coupled with observing and analyzing their usage patterns. This analysis provides insights into how a system like SmartMD can aid daily clinical work and stimulate further research and improvements.

Identify Challenges for Clinical Decision Support Systems. Fourth, we derive design challenges for CDSS based on our results and insights.

[1] `https://github.com/DATEXIS/smartmd-backend`

5.2.2. The Differential Diagnosis Process

We identify in the medical diagnosis process described by Altkorn, 2020; Croskerry, 2009 six data integration, retrieval, or prediction tasks for a DDx system (Figure 5.5). First, the physician needs to assess all available information regarding a patient, such as electronic health record entries, laboratory results, or medical images (section 5.2.2.1). Next, she needs to organize helpful resources for similar case data (section 5.2.2.2). Following this, she explores the data (section 5.2.2.3) and selects the most similar and relevant recorded cases (section 5.2.2.4). She has now formed a hypothesis regarding suitable clinical pathways and diagnoses for the patient. Now, she can additionally consult the most recent medical literature to narrow down, rank, and finally, choose the most probable clinical pathway for the patient (section 5.2.2.5). The corresponding clinical actions (section 5.2.2.6) follow this and lead to a clinical endpoint or new insights in the case at hand. Now she can start over the process with the new information.

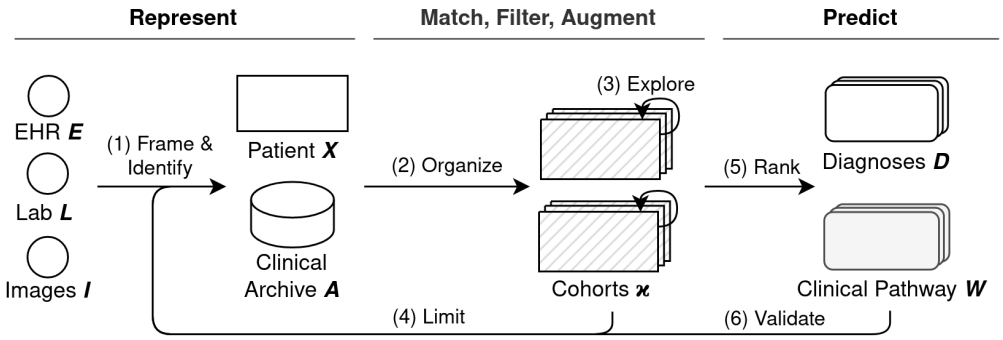


Figure 5.5.: The Differential Diagnosis Process. We condense the steps of the differential diagnosis into six steps. (1) Based on data from the patient's Electronic Health Record, we frame and identify the patient's current situation. Next (2), we model and organize similar cases from clinical archives into cohorts. Following this, the system takes feedback on the selected cohorts' relevance from the medical expert (3). She can now limit (4) the selection by adding or removing clinical concepts from the patients' situation representation X. SmartMD recommends (5) complementary or additional hypotheses regarding the diagnosis or clinical pathways. Finally, (6) the medical doctor can validate the recommendations and add new gained input to X.

5.2.2.1. Frame and Identify

In the first step, the medical doctor needs to assess all information available regarding the patient. She visits laboratory results, medical images , and prior assessment notes in the EHR. In doing so, she determines which medical concepts, described in medical ontologies, like *UMLS*, and coding systems, such as *ICD-10* or *SNOMED*, apply to the patient. Finally, she ends up with a set of concepts that describe the patient.

5.2.2.2. Organize

Next, the physician needs to retrieve records of patients that are similar to the patient at hand in the organize step (2). She does this by searching for patients that match the patient at hand in clinical concepts and demographics. Additionally, she might formulate further restrictions on the search based on her clinical experience and intuition. Now, she groups the retrieved patients into cohorts based on how well they fit the patient.

5.2.2.3. Explore

The explore step (3) allows the medical doctor to assess, compare, and drill down into the assembled cohorts. As a result of this, she can determine if a cohort is relevant to the differential diagnosis process.

5.2.2.4. Limit

She might add additional limitations to the query in the limit step (4). This refinement is necessary if she is not satisfied with the overall results of the "organize" step or spots interesting details while drilling down into patients.

5.2.2.5. Rank

The rank step (5) proposes complementary diagnoses or clinical pathways to the medical expert. She can now investigate multiple hypotheses and choose to take them into account when treating the patient. Since medical professionals know how to deal with a large variety of cases, a supporting system must aim to point

out rare but relevant diagnoses. Moreover, it must help her to take the best course of action to confirm a diagnosis and finally treat it.

5.2.2.6. Validate

Finally, she might decide to carry out the suggested clinical action. Doing so, she gains new insights, e.g., if a treatment is effective or diagnostic results are available. She can now update the system with new data, such as lab results, medical images, or clinical observations, and evaluate the new situation. This step influences all 'priors' of our model so that it can provide new insights.

5.2.3. Models and Methods

Our thoughtful analysis from the last section 'translates' core medical tasks into common data integration and machine learning tasks. In this section, we discuss important properties and formalize each task. Next, we integrate our recent work into an integrated system for DDx on the MIMIC-III dataset and the UMLS concept hierarchy. In particular, we focus on deep neural natural language processing models for clinical concept recognition and linking, negation detection, or to frame and identify a patient's situation. We model SmartMD using the microservice architecture paradigm (Sill, 2016) to scale up to tens of thousands of patients and deliver fast results. Finally, we visualize the results for doctors with Sankey diagrams.

5.2.3.1. Representing Patients from Clinical Text

Our focus is on free-form texts in EHRs. These texts contain information from the doctors' perception, which is often complementary to more structured data, such as vital- or lab data or other forms of diagnostics. Our interviews with clinical doctors and a deep manual analysis of many hundred clinical notes unrevealed at least three core tasks: clinical concept recognition, clinical concept linking, and negation detection for representing text in archives' records. Our models take advantage of deep learning models (Gu et al., 2021; A. Wang, Pruksachatkun, et al., 2019; A. Wang, Singh, et al., 2018) and transfer learning (Brown et al., 2020; J. Lee, Yoon, et al., 2019; M. E. Peters et al., 2018) to outperform in high

dimensional spaces. This is necessary since typical clinical notes often include typos, abbreviations, idiosyncratic language, other 'previously unseen' vocabulary, and bullet points instead of well-formed sentences (Leaman, Khare, and Lu, 2015).

Clinical Concept Recognition. Clinical Concept Recognition is the task of identifying words mentioning a clinical concept in a text and assigning those mentions to concept identifiers (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017; Si et al., 2019; Y. Wu, M. Jiang, et al., 2018). Deep Neural networks have shown to be effective in this task, especially in generalizing to examples not seen at training time (Arnold, Gers, et al., 2016; Y. Wu, M. Jiang, et al., 2018). For example, the methods of (Arnold, Gers, et al., 2016) and (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017) can adapt and scale to unseen inputs with just a few thousand training examples. We follow (Arnold, Gers, et al., 2016) and (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017) and formulate clinical concept recognition as a sequence tagging problem. We apply the widely used (Arnold, Gers, et al., 2016; B. Tang et al., 2013) *BIOES* tagging scheme, an extension of the BIO scheme (Ramshaw and Marcus, 1995). Thus, the model needs to assign the correct tag class out of $\{B, I, O, E, S\}$ to every token in a clinical note. We use sub-word embeddings (Bojanowski et al., 2017; Y. Wu, Schuster, et al., 2016) to ensure robustness when facing noise typical in clinical narratives such as misspellings, acronyms, or clinic-specific jargon (Chapman, Nadkarni, et al., 2011 Sep-Oct; Leaman, Khare, and Lu, 2015).

Training on MedMentions and Wikipedia. The MedMentions dataset (Mohan and D. Li, 2019) consists of 4,392 PubMed abstracts annotated by professional annotators with UMLS concepts. We chose the ST21pv training subset for training our model. The ST21pv subset covers less but clinically relevant UMLS concepts (25,419) than the full MedMentions dataset (34,724). Heilman and West, 2015 survey the quality of Wikipedia's medical section and conclude that it is a sound source on medical topics. Therefore, we use the medical part of Wikipedia as an additional training resource as originally proposed by (Bunescu and Paşca, 2006). Consequently, we download a dump of Wikipedia and filter it for pages on medical topics. Next, we query Wikidata (Vrandečić and Krötzsch, 2014) for medical concepts such as diseases or medications and all available synonyms or

alternative names. Following this, we use the work of Mandel, 2019 to annotate additional relevant medical concept mentions in the Wikipedia dump.

Clinical Concept Linker. This task normalizes each medical concept from the recognizer and assigns a unique identifier from a controlled vocabulary or ontology (Fu et al., 2020; Tseytlin et al., 2016). Covering most internationally used medical coding systems, the unified medical language system (UMLS) (Bodenreider, 2004) serves as our ontology. Every concept has a Concept Unique Identifier (CUI) associated with multiple names or mentions in the UMLS. For example, *common cold* and *head cold* refer to the same CUI *C0009443*.

We index all 10,406,797 English concept names and their 4,413,092 concept identifiers. Therefore, we rely on full-text indexes provided by Elasticsearch (*Elasticsearch* 2021). For every English concept name in the UMLS, we create an entry consisting of its CUI, name, description, and semantic type. We analyze the name and the description using tokenization (Webster and Kit, 1992) and stopword removal. Additionally, we perform 3-gram-subword-tokenization (Bojanowski et al., 2017) for the concept names to be robust for spelling errors. Furthermore, this pre-processing allows serving an interactive search-as-you-type scenario proposed by Arnold, Dziuba, and Löser, 2016 and used in Section 5.2.3.4.

$$\begin{aligned} \text{link}(C, M) &= \text{rescore}(\text{BM25}(C, m_{ngram}), d) \\ \text{rescore}(C, d) &= \gamma \cdot (\text{BM25}(C, d) + \text{BM25}(C, m)) \end{aligned} \tag{5.3}$$

Next, we model clinical concept linking similar to (Schumacher, Mulyar, and Dredze, 2020) as a ranking scenario. Likewise, the linker needs to rank all candidate concepts C_{cui} given a mention M, which consists of the text of the mention m and its context d.

$$\begin{aligned} \text{BM25}(C, Q) &= \sum_{i=1}^{n} \text{IDF}(q_i) \frac{f(q_i, C)(k_1 + 1)}{f(q_i, C) + k_1 \left(1 - b + b \frac{|C|}{\text{avgcl}}\right)} \\ \text{IDF}(q_i) &= \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \end{aligned} \tag{5.4}$$

We employ BM25 ranking (5.4) with character-n-gram matching and context-based reranking (Rajani, Bornea, and Barker, 2017) in Equation 5.3, where γ is a

scaling factor. We use the top one candidate as the predicted concept.

UMLS Semantic Type	→	SmartMD Medical Type
Clinical Drug		Medication
Pharmacologic Substance		Medication
Antibiotic		Medication
Sign or Symptom		Symptom
Laboratory or Test Result		Symptom
Clinical Attribute		Symptom
Hazardous or Poisonous Substance		Symptom
Disease or Syndrome		Symptom
Bacterium		Symptom
Virus		Symptom
Fungus		Symptom
Eukaryote		Symptom
Finding		Symptom
Mental or Behavioral Dysfunction		Symptom
Neoplastic Process		Symptom
Cell or Molecular Dysfunction		Symptom
Injury or Poisoning		Symptom
Therapeutic or Preventive Procedure		Therapy
Diagnostic Procedure		Diagnostic
Laboratory Procedure		Diagnostic

Table 5.3.: Mapping of relevant UMLS semantic type names to clinical types used in SmartMD. We assign any UMLS semantic type names not contained in this table to the "*Other*" class.

Clinical Type Classification. The linked concepts refer to a semantic type within the UMLS. The UMLS organizes these fine-grained types in a hierarchy containing 91 types. To support medical professionals' focus on the differential diagnosis process, we condense this typing system down into five classes: "Medication," "Symptom," "Diagnostic," "Therapy," and "Other." Accordingly, we use a handcrafted dictionary for this normalization step, as shown in Table 5.3. As a result, our clinical concept linker yields triples of the form (mention, cui, medical type).

Assertion Detection with fine-tuned BERT. Clinical narratives often contain words modifying the *presence*, *absence*, or *possibility* of clinical concepts. For example, a report in an EHR might state, "*The patient denied any shortness of breath.*" Consequently, our system must mark "*shortness of breath*" as *absent* in the patient representation.

Especially, detecting the often vaguely expressed "possible" class is challenging. van Aken, Papaioannou, et al., 2021 discuss the helpfulness of pre-trained language models to tackle this problem. They show that BERT-based models significantly increased scores in the often underrepresented "possible" class of *0.786 F1*. Therefore, we follow their recommendation to use the BERT-based (Vaswani et al., 2017) approach of Alsentzer et al., 2019. They solve the task of entity-specific Assertion Detection by fine-tuning (Howard and Ruder, 2018) BERT on biomedical paper abstracts and discharge summaries. Consequently, we can profit from BERTs' vast language understanding and generalization capabilities.

$$\begin{aligned} \mathrm{G}(\alpha_i \mid \mathrm{BERT}, \theta) &= \mathrm{softmax}\left(\mathrm{BERT} \cdot V^T\right) \\ &= \frac{\exp\left(P\left(\alpha_i \mid \mathrm{BERT}, \theta\right)\right)}{\sum_{j=1}^{\alpha} \exp\left(P\left(\alpha_j \mid \mathrm{BERT}, \theta\right)\right)} \end{aligned} \tag{5.5}$$

Following Alsentzer et al., 2019; van Aken, Trajanovska, et al., 2021, we model assertion detection as a three-class classification problem where we aim to predict the likelihood of each of the three classes present, absent, or possible. Accordingly, we approximate the probability of each class ($\alpha = \{\text{present}, \text{absent}, \text{possible}\}$) given the BERT model and all trainable parameters θ as illustrated in Equation 5.5. To do so, we apply the fine-tuned BERT model on the input vector V and calculate the final result using the softmax function. V is a sequence of words surrounding the concept mention in the clinical narrative. Finally, we use the class with the highest probability and assign it to the corresponding clinical concept C.

5.2.3.2. Cohort Retrieval and Clustering

Selecting patients for cohort identification is an elementary task in clinical research. For this purpose, it is necessary to define the similarity of patients to model helpful cohorts. We focus our cohort modeling method on patient

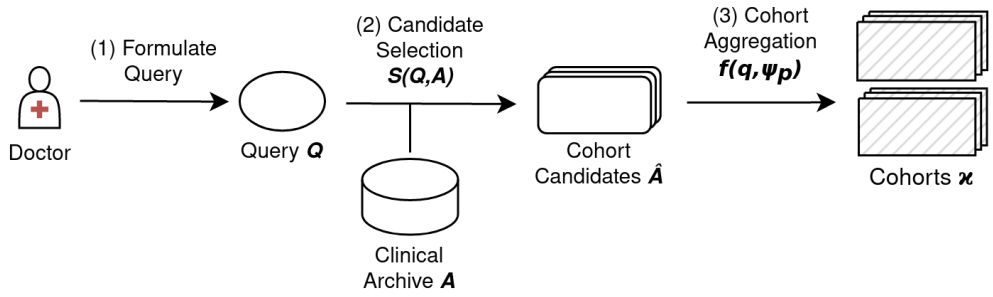


Figure 5.6.: Cohort Modeling Process in SmartMD. (1) A medical doctor formulates a query that describes the situation of a patient and her first assessment of the case. (2) Next, the system retrieves candidates $\hat{A}$ from the clinical archive A using scoring function $S(Q,A)$. (3) Following this, SmartMD aggregates the candidates into cohorts κ using our aggregation function $f(q,\psi_p)$.

characteristics derived from free-text reports, similar to Glicksberg, Miotto, et al., 2018; Glicksberg, Oskotsky, et al., 2019; J. Lee, Maslove, and Dubin, 2015; Sarmiento and Dernoncourt, 2016. Following J. Sun et al., 2012, we use medical experts' knowledge in our model. In particular, the SmartMD system enables medical professionals to choose the best-fitting features extracted from text to form a query against the patient archive. Our interviews with doctors revealed several types of results from a cohort search. First, doctors desire cohorts that ideally match their query. Moreover, doctors are interested in additional cohorts that share important query predicates and share common anomalies not requested explicitly in the query. As an illustration, assume that a medical doctor queries SmartMD for our running example. This results in the best matching cohort matching all query criteria; second, a cohort with no renal transplantation; third, another cohort with patients who commonly suffer from diarrhea. In the following and Figure 5.6, we describe candidate retrieval and cohort clustering.

Candidate Retrieval & Query Types. SmartMD supports boolean conjunctions of predicates, including polarities and negations, such as "*include all patients with Type 2 Diabetes, who have undergone renal transplantation and exclude all patients with present Diarrhea.*" Figure 5.7 provides an overview of possible expressions. The system permits clinical archive A with selected concept names $\psi = \{c_1, ..., c_i\}$ to obtain candidates $\hat{A}$. Additional query parameters consist of predicates that

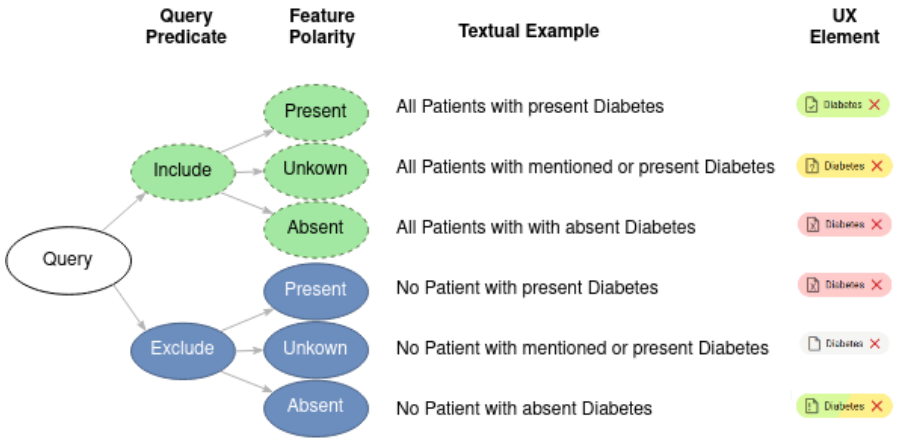


Figure 5.7.: This figure shows possible query predicates supported by the cohort modeling service of SmartMD. A query predicate can either be of the query predicate type include or exclude. Combining this with the feature polarity results in query predicates that express the shown textual examples, e.g., for diabetes.

either include $\rho_j = 1$ or exclude $\rho_j = -1$ a patients' concepts. Consequently, Q is a set of the tuples of all included or excluded concepts $Q = \{(\phi_1, \rho_1), ..., (\phi_n, \rho_n)\}$ and $\Phi = \{\phi_1, ..., \phi_M\}$ is the set of all concept names in Q.

$$S(x,Q) = \sum_{j}^{|\Phi|} \sum_{i}^{|\Psi|} \Big(\text{BM25}(c_i, \phi_j) \cdot \max(v_i \cdot \rho_j, 0) \Big)$$
$$\hat{A} = \{x | x \in A \text{ and } S(x,Q) > T\} \tag{5.6}$$

We index the electronic health records together with their features from our patient representation step. Next, we apply the BM25-based scoring function S to find similar patients based on the concept names in Equation 5.6. To consider negations, we filter candidates' concepts by calculating the maximum between the product of its polarity v_i, and the queried concepts predicate ρ_j and zero. This filter removes the BM25 score of every concept that does not match the queries' required polarity from the candidate's score. Subsequently, we use function S to build a set of cohort candidates $\hat{A} = \{\hat{x_1}, ..., \hat{x_p}\}$ that surpass the threshold T. Finally, we use $\hat{A}$ as input for the cohort aggregation step.

Cohort Clustering. We cluster the cohort candidates $\hat{A}$ into the final cohorts κ. Remember, we seek one best cluster of patients with best-fitting query predicates and additional clusters that match parts of the query but share other commonalities. Our aggregation function in equation 5.7 organizes retrieved patient candidates into the best-fitting cohort.

$$\begin{aligned} \kappa_l &= argmax_{q \in \Omega(\Phi)}(f(q, \psi_p)) \\ f(q, \psi_p) &= \frac{exp(J(q, \psi_p))}{\sum_{q \in \Omega(\Phi)} exp(J(q, \psi_p))} \end{aligned} \tag{5.7}$$

We calculate the probability that a candidate at index p belongs to a cohort q for every possible cohort. For this purpose, we define the set of all cohorts as the power set of all concept names in the query ($\kappa = \Omega(\Phi)$). Next, we approximate the probability of exclusive cohort membership for each candidate by calculating the argmax of the scoring function f and the softmax of the patient similarity function J to obtain a distinctive score.

Candidate Similarity. We model the similarity between retrieved patient candidates and a cohort as the distance function J. It takes the candidate's set of medical concepts ψ_p and concepts of the current cohort q as arguments. Subsequently, we employ the Jaccard distance (Levandowsky and D. Winter, 1971) in Equation 5.8 to estimate the candidate's similarity.

$$J(q, \psi_p) = \frac{|q \cup \psi_p| - |q \cap \psi_p|}{|q \cup \psi_p|} \tag{5.8}$$

In essence, we model the probability that a candidate belongs to a cohort by the number of medical concepts that match the cohort. This method is easily transferable to determine the similarity between patients.

5.2.3.3. Predicting Clinical Actions

A complex web of interactions underlies the associations between diseases. This is particularly the case for the modern diseases of the developed world (Arandjelović,

Correlation Type	Medical Type	
	Source	Target
Symptoms-Treatment	Symptom	Therapy, Medication
Symptoms-Diagnostics	Symptom	Diagnostic
Diagnosis	Symptom	Symptom

Table 5.4.: Correlation types supported by SmartMD

2015). Observing, analyzing, and understanding these dependencies between medical concepts in cohorts is a common method in medical research. For example, Arandjelović, 2015 and Landi et al., 2020 analyze past hospital admissions. We follow this idea and enable SmartMD to recommend clinical actions and propose diagnoses based on correlations observed in the hospital archives data. Therefore, we take the previously selected cohorts of historical cases into account and enable clinicians to perform such analyses in their day-to-day work.

It is crucial for such a system to carefully model the relevance of clinical actions such as diagnostics, treatments, or diagnoses and minimize the number of irrelevant suggestions to win the users' trust (R. T. Sutton et al., 2020). Therefore, we follow Watford et al., 2018, Miotto, F. Wang, et al., 2018 and Landi et al., 2020 and base, as in our previous work (Schneider, Arnold, et al., 2018) the recommendations of SmartMD on correlations between medical concepts observed in expert-selected documents such as EHRs. Thereby, we differentiate between the three correlation types: Symptoms-Treatment, Symptoms-Diagnostics, and "Diagnosis." Table 5.4 gives a detailed overview of them and their respective medical types. Resultantly, we define a correlation type CT as a tuple of medical source M_s and target M_t types. Thus, we formulate the clinical action recommendation as a distribution of correlation values between the source and target medical types of a given type with respect to the selected cohorts κ.

Ranking Clinical Actions. We use a relational database[1] to enable cohort-based clinical action recommendation at scale. Thereby, we can support recommendations based on cohorts of thousands of relevant patients and their respective clinical concepts. Accordingly, we transform the encoded EHR data into a rela-

[1]We use PostgreSQL 12 `https://www.postgresql.org/`

tional database schema (Codd, 1970). Thus, we arrange all clinical concepts in relation F as shown in Equation 5.9 and group every concept's name and polarity by the document id. As a result, we obtain distinct tuples of concept name and polarity.

$$F(\text{id}, \text{doc\_id}, \text{text}, \text{polarity}, \text{type}) \tag{5.9}$$

Filtering Clinical Concepts. Next, we query the database for clinical concepts that belong to the patients in the selected cohorts κ. We filter the two projections of F, F_1, and F_2 for the queries' correlation type by matching the concept's medical type (Equation 5.10). We filter F_1 for M_s and F_2 for M_t. We exclude all clinical concepts that are stated as absent since we want to recommend only relevant concepts.

$$\begin{aligned} F_1 &:= \sigma_{\substack{\text{polarity}\neq\text{ABSENT} \\ \wedge\text{type}=M_s \\ \wedge\text{d\_id}\in\kappa}}(F) \\ F_2 &:= \sigma_{\substack{\text{polarity}\neq\text{ABSENT} \\ \wedge\text{type}=M_t \\ \wedge\text{d\_id}\in\kappa}}(F) \end{aligned} \tag{5.10}$$

Approximating Probability of Concepts in Cohorts. As the next step, we count and group the occurrence of every remaining clinical concept in Equation 5.11. We use the concept names as grouping keys. Following, we divide the counts by the number of patients in the selected cohorts $|\kappa|$. As a result, we obtain the relations P_s and P_t, which approximate the likelihood of the respective clinical concepts. Eventually, we join the probabilities P_s and P_t with the remaining data in the relations F_1 and F_2 by using the concept name as the join condition.

$$\begin{aligned} P_s &:= \text{text}\,\mathrm{G}_{\frac{\text{count(text)}}{|\kappa|}}(F_1) \\ F_1 &:= F_1 \bowtie_{F_1.\text{text}=P_s.\text{text}} P_s \\ P_t &:= \text{text}\,\mathrm{G}_{\frac{\text{count(text)}}{|\kappa|}}(F_2) \\ F_2 &:= F_2 \bowtie_{F_2.\text{text}=P_t.\text{text}} P_t \end{aligned} \tag{5.11}$$

Next, we join in Equation 5.12, the source concepts relation F_1 with the target concepts relation F_2 based on the respective Patient. As a result, we obtain every existing combination of co-occurring clinical concepts F_{1_2} within the selected cohorts κ coupled with their single probability.

$$F_{1\_2} := F_1 \bowtie_{F_1.d\_id=F_2.d\_id} F_2 \tag{5.12}$$

As the next step, we obtain F_{1_2B} as a projection of F_{1_2}. Similar to the single probabilities P_s and P_t, we calculate the joint probability of their co-occurrences $P_{s,t}$ in Equation 5.13. Additionally, we compute their absolute occurrences. Lastly, we join $P_{s,t}$ to F_{1_2} on the source and target concept names so that we end up with the single probabilities and the probability of every co-occurring pair of source and target concepts in relation F_{1_2}.

$$\begin{aligned} F_{1\_2B} &:= \Pi_{\text{text}_{F_1},\text{text}_{F_2}}(F_{1\_2}) \\ P_{s,t} &:= \text{text}_{F_1}\,\text{text}_{F_2}\,\mathrm{G}_{\frac{\text{count}(\text{text}_{F_1},\text{text}_{F_2})}{|\kappa|},\text{count}(\text{text}_{F_1},\text{text}_{F_2})}(F_{1\_2B}) \\ F_{1\_2} &:= F_{1\_2} \bowtie_{\substack{F_{1\_2}.\text{text}_{F_1}=P_{x,y}.\text{text}_{F_1} \\ \wedge F_{1\_2}.\text{text}_{F_2}=P_{x,y}.\text{text}_{F_2}}} P_{s,t} \end{aligned} \tag{5.13}$$

Scoring Clinical Concepts. Once we obtain all co-occurring clinical concepts in our selected cohorts, we need to model their relevance. We follow Watford et al., 2018 and the approach of our previous work (Schneider, Arnold, et al., 2018) and use point-wise mutual information (PMI) to measure the association of the observed co-occurrences. (Equation 5.14).

$$\text{pmi}(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{5.14}$$

For better interpretability, we follow Bouma, 2009 and divide the PMI by the informed self-information $h(x,y)$ in Equation 5.15 to introduce upper- (1) and lower-bounds (-1) to the measure. This results in the property that a value of 1 can be interpreted as a perfect correlation, 0 as independence, and a value of -1

as a contradiction.

$$\begin{aligned} \text{npmi}(x;y) &\equiv \text{npmi}(P(x,y),P(x),P(y)) \\ &= \frac{\text{pmi}(x;y)}{h(x,y)} \\ &= \left(\log \frac{p(x,y)}{p(x)p(y)}\right) \Big/ -\log p(x,y) \end{aligned} \tag{5.15}$$

A flaw of NPMI is its bias towards low-frequency co-occurrences, resulting in unrealistically high association values for rare concept combinations (Bouma, 2009; Watford et al., 2018). Worse, we observe that this causes upstream extraction errors and other noise to receive high scores. Consequently, we filter all recommendations by their absolute number of occurrences with the threshold T to counteract this problem. Empirically,. we have determined that a value of $T = 3$ serves good results without discarding interesting rare concept combinations.

$$\Pi_{\text{text}_{F1},\text{text}_{F2},\text{npmi}(P_{x,y},P_x,P_y)}(\theta_{\substack{\text{npmi}\geq 0 \\ \wedge\, \text{count}>T}}(F_{1\_2})) \tag{5.16}$$

Next, we calculate the NPMI for every pair of source and target clinical concepts. Remember, our goal is to recommend clinical actions. Hence, we filter the results to contain only values with an NPMI value of at least zero. Identically, we apply the Threshold T as a filter. After these filtering steps, we calculate the final projection containing the source and target concept names and their NPMI in Equation 5.16.

5.2.3.4. Interactive Feedback from Medical Experts

SmartMD offers easy-to-use interfaces to our previously defined clinical operators. We aim to ease the day-to-day work of medical professionals. Therefore, SmartMD offers a web-based front-end supported by most modern web browsers. To offer interfaces with low visual complexity (Miniukovich, Sulpizio, and De Angeli, 2018), we follow the design guidelines of material design[1]. Following Eiband et al., 2018 and Dudley and Kristensson, 2018, we aim for transparency of the

[1] `https://material.io/`

deep learning and statistical methods. Consequently, SmartMD provides multiple feedback loops and high-level abstractions of the system's state to the user.

Modeling the Index Patient. We enable the user to provide feedback to the system at any time. For example, when analyzing a patient's clinical notes, the Smart MD: DDx offers a top-down view of the found and linked concepts, as shown in Figure 5.8. The medical professional can now select which detected clinical concepts are relevant for further investigation in the differential diagnosis process. If the system misses a concept or a new relevant comes to the physician's mind, she can enter it in an auto-completed text field. The medical professional can assign the clinical features to one of the query predicates *Present*, *Possible*, *Present / Possible*, *Excluded*, or *Not Mentioned*. We visualize these predicates using color coding and icons from the iconify[1] collection. Moreover, the physician can determine if a clinical feature is absent or present in the patient at hand. SmartMD uses this input to build a query according to the query type definitions in section 5.2.3.2 and Figure 5.7.

Interactive Cohort Modeling. SmartMD shows in the next step the cohort modeling interface, see Figure 5.6. We use a data table[2] to show high-level information on the found cohorts, like how well they match the query or which points differ. The physician can drill down into the cohorts and inspect the EHR of every patient. This enables medical professionals to perform cohort analysis at scale and in their day-to-day work. Consequently, she can now include cohorts for further analysis or go back to the patient modeling step to refine her query.

Explaining Complimentary Recommendations. To visualize the results of the correlation analysis, we use Sankey diagrams[3] (Figure 5.10) as Ronicke et al., 2019 have shown their effectiveness in comprehending medical causalities and correlations. The diagram shows the relationship between, for example, symptoms and diagnostic measures. The size of the connecting line determines how much a

[1] `https://iconify.design/`
[2] `https://material.io/components/data-tables`
[3] `https://github.com/d3/d3-sankey`

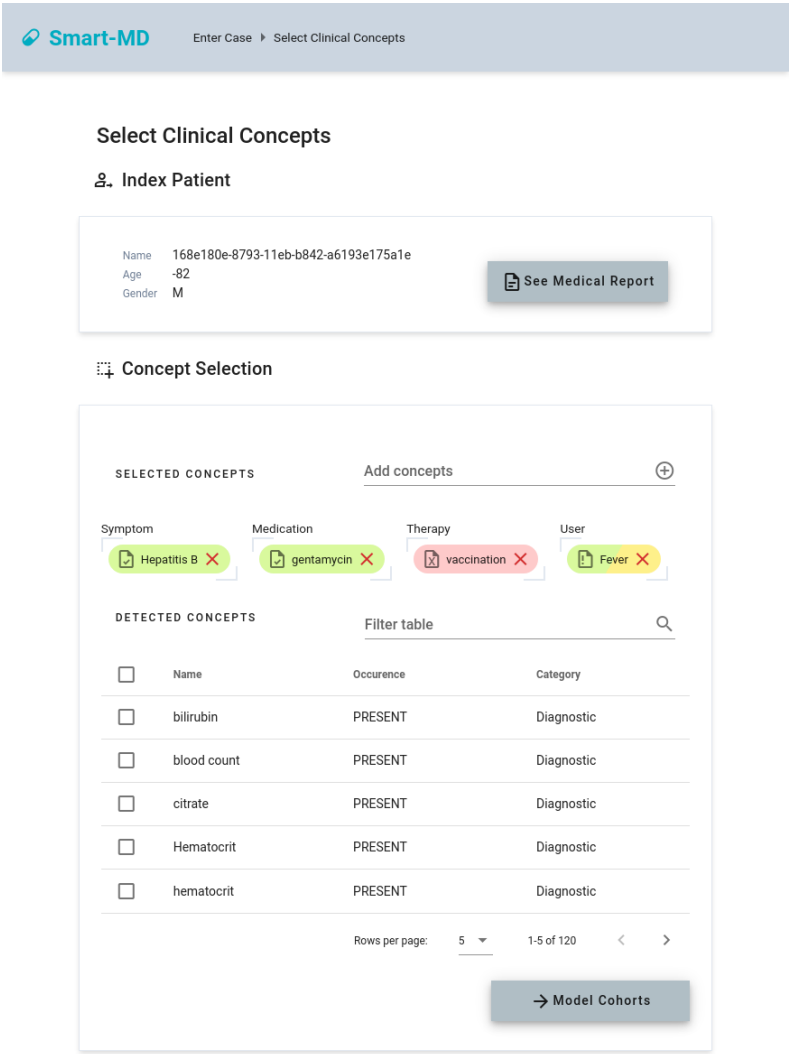


Figure 5.8.: Screenshot of the patient modeling view. SmartMD provides a high-level overview of the found concepts and information regarding the case at hand. This view lists all patient features and offers the ability to filter them. The middle part of the view is the filter query area. All selected features that should contribute to the cohort modeling process are listed and color-coded about the applied query predicate. SmartMD offers an additional auto-complete field to filter queries based on patient features that are not present in the case at hand.

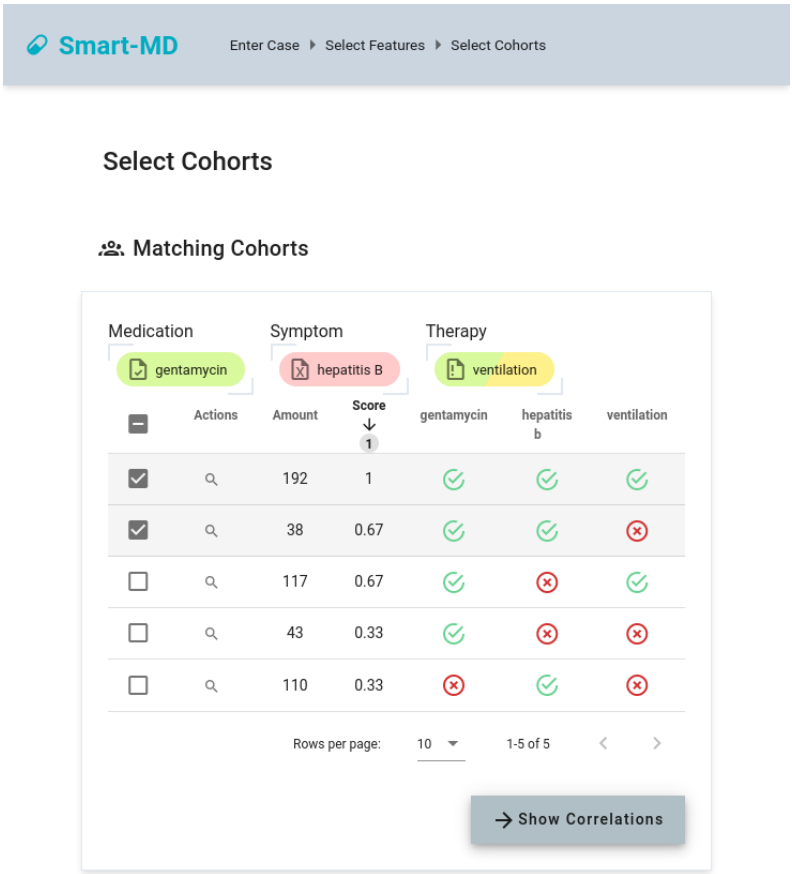


Figure 5.9.: Screenshot of the cohort modeling interface. We organize the retrieved candidate cohorts in a sortable data table. This table enables the medical professional to select cohorts relevant to the clinical action proposal step. Along with that, the color and icon coded difference between the query and any candidate cohort are made transparent. To reduce the cognitive load of the user, we additionally show the executed query.

symptom and a diagnostic action are associated. Additionally, we provide the NPMI score and the absolute co-occurrence values while mousing over a connecting line.

Transparently Recommending Clinical Actions at Scale. Considering all these steps, SmartMD enables clinicians to obtain complimentary recommendations for clinical actions and diagnoses based on in-depth cohort analyses from EHR text data at scale. The system explains its recommendations transparently in all steps

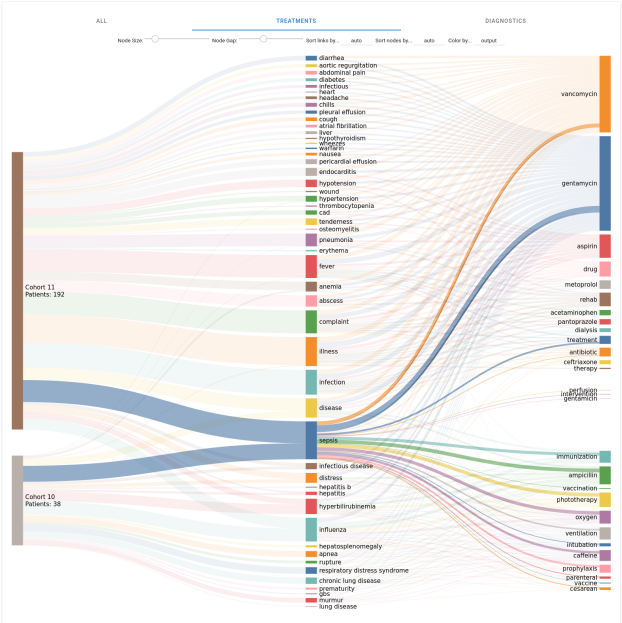


Figure 5.10.: Screenshot of the correlation visualization. We use a Sankey diagram to visualize co-occurring clinical concepts. This visualization supports three modes for every supported correlation type (Table 5.4).

by providing easy access to the plain text EHR and visualizing the used data.

5.2.4. Evaluation

We evaluate the differential diagnosis process implemented by SmartMD in two detailed steps. We ask medical experts to use the system to assess two cases in the MIMIC-III (Johnson et al., 2016) dataset. We collect observational data during this experiment and perform a structured interview with the medical experts afterward.

5.2.4.1. The MIMIC-III Dataset

MIMIC-III is an electronic health record dataset that is publicly available for researchers. The data describes ICU patients at Beth Israel Deaconess Medical Center in Boston, US. We chose MIMIC III for our evaluation for two reasons. First, there is a lack of publicly available data that can be used in research and permits

Property	Value
Patients	33.204
ICU visits	44.026
Unique diseases	6.184
Diseases with only 1 occurrence	1.355
Total disease mentions	521.786
Unique medications	3.725
Medications with only 1 occurrence	315
Total medication mentions	1.885.865
Unique procedures	1.947
Procedures with only 1 occurrence	366
Total procedure mentions	210.376

Table 5.5.: Statistical analysis of the MIMIC-III data set.

the reproduction of experiments. Second, is the large amount of data generated at ICUs and the necessity for proactive, precise, and personalized care strong indicators that the processes here could vastly profit from computerized CDSS (Sharafoddini, Dubin, and J. Lee, 2017).

MIMIC-III consists of 53,423 distinct admission records between 2001 and 2012 that contain free-text reports and structured data. We aim with SmartMD to provide a generally applicable decision support system, which is also suitable to other medical domains that profit from digital decision support systems, e.g., chronic disease treatment (Cui, Bozorgi, et al., 2012; Sharafoddini, Dubin, and J. Lee, 2017). For this reason, we focus our evaluation on the free text reports, which are often less incomplete in non-ICU settings (Sarmiento and Dernoncourt, 2016). For illustration purposes, we create an artificial example case similar to MIMIC-III:

> The patient is a [age] year old gentleman with a history of hypertension and known aortic aneurysm. He was being followed by [** medical professional 123 **]. He presented to [**Hospital 1 **] with worsening abdominal and back pain. Pain began about a day prior to admission and gradually worsened. It is not associated with food. No vomiting, no chest pain, shortness of breath. At [**Hospital 2 **], the

patient underwent abdominal CT scan which did not show a ruptured aneurysm. The patient was transferred to [**Hospital 1 **] for further management.

MEDICAL HISTORY: Significant for hypertension; aortic aneurysm; gout;.

MEDICATION ON ADMISSION:

ALLERGIES: Amoxicillin

PHYSICAL EXAM:

FAMILY HISTORY:

SOCIAL HISTORY:

Additionally, table 5.5 provides statistical insights on the data.

We randomly draw 2 cases from the MIMIC-III-based dataset of van Aken, Papaioannou, et al., 2021 for manual evaluation involving medical experts for our experiments.

5.2.4.2. Qualitative Evaluation

To verify our approach's effectiveness, we ask five medical experts (Table 5.7) to work on two randomly drawn cases and recommend treatments, diagnostics, and diagnoses that apply to the cases. Accordingly, we present the experts similar to the evaluation of van Aken, Papaioannou, et al., 2021 each case's admission notes. They can now use the SmartMD system to work on the cases. Next, we invite them for a structured interview. Correspondingly, we formulate hypotheses and related interview questions. Finally, we analyze and interpret the answers.

Hypotheses & Interview Questions. We aim to assess the correctness, helpfulness, and complementariness of our system and process. Moreover, we aim to discover challenges in designing and implementing the deep learning-aided differential diagnosis process. For this motivation, we formulate the following hypotheses:

- H1 SmartMD delivers complementary medical hypotheses

- H2 SmartMD's recommendations are medically sound.
- H3 SmartMD is helpful for medical professionals.
- H4 SmartMD's recommendations are comprehensible.
- H5 SmartMD explains how it arrives at conclusions.

We categorize and group the interview questions in Table 5.6. We sort questions by each hypothesis that we want to survey. Moreover, we assign "open" or "range" to every question as a question type. Following Likert, 1932, we ask the participants how much they agree with the statements of range-type questions on a scale from 1 (Strongly disagree) to 5 (Strongly agree).

Id	Question / Statement	Type	Range
P	**Preface**		
P1	What is your highest medical-related degree?	open	n/a
P2	How many times have you used SmartMD? (single occasions, e.g., 0-3)	open	n/a
P3	How old are you?	open	n/a
H1	**SmartMD delivers complementary medical hypotheses.**		
H1.1	The system recommended a treatment that I hadn't thought of beforehand.	range	1 - 5
H1.2	The system recommended a diagnostic that I hadn't thought of beforehand.	range	1 - 5
H1.3	The system recommended a diagnosis that I hadn't thought of beforehand.	range	1 - 5
H2	**SmartMD's recommendations are medically sound.**		
H2.1	The proposed treatments, diagnoses, and diagnostic procedures are medically sound.	range	1 - 5
H2.2	The predicted diagnoses were correct.	range	1 - 5
H3	**SmartMD is helpful for medical professionals.**		
H3.1	SmartMD is helpful to get an overview of similar cases.	range	1 - 5
H3.2	The recommended treatments were helpful.	range	1 - 5
H3.3	The recommended diagnostics were helpful.	range	1 - 5
H3.4	I would be interested in working with the system in the future.	range	1 - 5
H4	**SmartMD's recommendations are comprehensible.**		
H4.1	I had trouble comprehending the system's recommendations.	range	1 - 5
H4.2	The system's reasoning for its recommendations was comprehensible.	range	1 - 5
H5	**SmartMD explains how it arrives at conclusions.**		
H5.1	I was able to retrace the system's decision-making process.	range	1 - 5
H5.2	The decision-making process was transparent and understandable.	range	1 - 5

Table 5.6.: Interview questions and hypotheses

Participant	#1	#2	#3	#4	#5
P1	Staatsexamen	Dr. med.	MD	MD	PD
P2	2	3	1	1	1
P3	32	27	29	28	30

Table 5.7.: Self-reported demographic features of experiment participants.

Results. Table 5.7 presents self-reported demographics based on our preface questions P1 - P3. We observe that the participants' age ranges from 27 to 32 years. The medical degrees range from entry-level ("Staatsexamen") to professorship level (PD - "Privatdozent"). Two participants reported that they used earlier versions of SmartMD before the experiment.

We report the results of the hypothesis-oriented questions H1.1 - H5.2 in Table 5.8. We group the results of each question by its associated hypothesis. Likewise, we group reported agreement values into three categories: negative (values < 3), neural (values $= 3$), and positive (values > 3). We observe the following results, which section 5.2.5 discusses in detail:

- Hypothesis 1: two positives, two neutral, and 11 negative values of 15 reported values.
- Hypothesis 2: no positives, two neutral, and eight negative values of 10 reported values.
- Hypothesis 3: four positives, seven neutral, and nine negative values of 20 reported values.
- Hypothesis 4: three positives, six neutral, and one negative value of 10 reported values. Please note that question H4.1 asks for a negative experience. Accordingly, we interpret reported values < 3 as positive and > 3 as negative results for this question.
- Hypothesis 5: two positives, five neutral, and three negative values of 10 reported values.

Question Id / Participant	#1	#2	#3	#4	#5
H1.1	1	2	2	2	4
H1.2	4	2	2	2	2
H1.3	2	3	2	3	2
H2.1	3	3	1	2	2
H2.2	2	2	2	2	2
H3.1	1	4	2	3	4
H3.2	2	3	1	3	2
H3.3	3	3	2	3	2
H3.4	3	2	2	4	4
H4.1	5	2	3	3	1
H4.2	3	3	3	3	4
H5.1	2	3	3	4	3
H5.2	2	3	2	3	4

Table 5.8.: Results of the structured interview.

5.2.4.3. User Observation Study

Understanding medical experts' "information needs" is crucial in designing clinical decision support systems and processes. We designed and implemented the deep learning-aided differential diagnosis process based on preliminary observations collected using our previous work in (Schneider, Arnold, et al., 2018). Our goal is to verify and refine our assumptions that lead to the presented DDx process.

Therefore, we collect usage data during our experiments and record every click taken. Specifically, we record interaction events for every worked case in the experiment. These events contain a session identifier, the current URL, the Referring URL, and the used UI element's identifier. The resulting dataset allows us to model the user's paths and the differential diagnosis process through the system. For this purpose, we assign and categorize each observable event to a step in the differential diagnosis process described in section 5.2.2, and Figure 5.5.

Results. Collecting interaction events during our experiments enables us to aggregate single events across participants and visualize this data in a user journey map. Figure 5.11 shows all data collected during our experiment. We present all

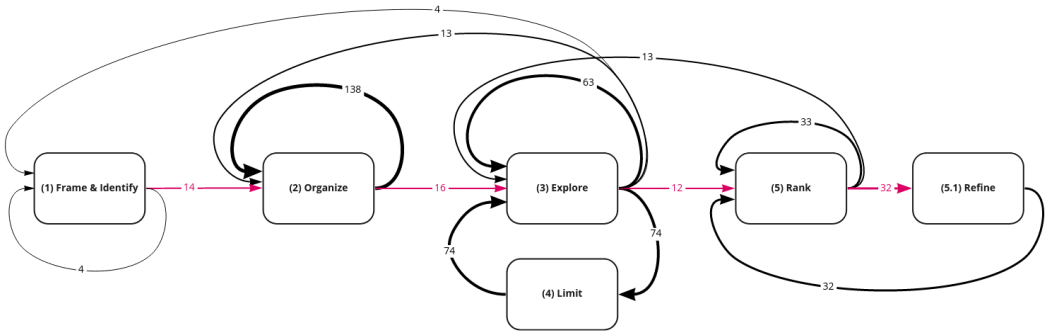


Figure 5.11.: User journey in the SmartMD Prototype. This diagram shows the path taken by the medical experts through the differential diagnosis process when working on the cases of our evaluation.

actions taken in the system but do not visualize the endpoints of single sessions. Thus, the numbers attached to process step transitions might not add up since the participant might interrupt and return the experiment session at any time or use their web browser's functions to navigate the application.

We indicate the main process flow using red color, and it follows the intended flow as suggested in section 5.2.2. We observe that the participants make just a few iterations in the Frame & Identify phase and move forward to Organize quite fast. We could only observe four self-reflexive iterations for Frame & Identify, which is a small number of steps compared to 138 iterations in the Organize phase. Likewise to the organize phase, we observe many (137) iterations within the Explore and Limit phases. Contrary to phases 2, 3, and 4, the number of self-reflexive iterations decreases for steps Rank and Refine.

Interestingly, in practice, we do not observe iterations that span the whole process, going back from step 5 to step 1. Likewise, we register not a single iteration back from the rank phase to frame and identify. Different from this observation, we see 13 instances where participants looped back from the ranking phase to explore most likely to refine the selection criteria for cohorts. Moreover, we observe four actions of moving from the explore phase back to frame and identify. In 13 cases, the participants moved from the explore stage back to organize.

5.2.5. Discussion

In this section, we discuss our results, present our findings, and describe design challenges for text-based CDSS.

5.2.5.1. Findings

SmartMD occasionally Recommends Complementary Treatments and Diagnostics. We report that two participants received complimentary recommendations from the SmartMD system. Two other participants report a neutral sentiment regarding recommended diagnoses. Some participants observed unhelpful recommendations. Overall, we conclude that HYPOTHESIS H1 requires further research to explore error sources and improve the helpfulness of the system.

Additional Medical Data is Required. Our experiment setup is limited to the MIMIC-III dataset as a single data source. Thus, the system needs to extract all medical knowledge from this rather small data set. Correspondingly, participants reported low scores regarding medical soundness. Thus, we need to reject HYPOTHESIS H2 and conclude that SmartMDs recommendations are not medically sound and need improvement. The system lacks information stored in medical knowledge bases, like UMLS, research papers, and medical textbooks. Moreover, the data in MIMIC-III is biased by its nature of covering ICU cases exclusively. Incorporating these multi-modal data sources poses an interesting challenge for future work.

SmartMD can be Helpful to Obtain an Overview of Similar Cases. Two participants reported that SmartMD was helpful for them to get an overview of similar cases. Three participants would like to work with SmartMD in the future.

SmartMDs Recommendations are Comprehensible. Most participants reported that the recommendations of SmartMD were comprehensible, and they were often able to understand systems decisions. In conclusion, we confirm HYPOTHESIS H4 and H5.

Experts are committed to the selected index patient once cohorts are selected. We observe that medical doctors perform many refinement and exploration steps during our experiments, sometimes cycling back to earlier process stages. (See Figure 5.11) Contrary to this, we observe no direct cycle back to Frame & Identify or Organize. Alongside this, we see that most of the refinement actions take place in the Organize, Explore, and Limit phases. Moreover, we see here more steps back to earlier process steps.

Better Medical Language Understanding Models are Required. We observe shortcomings in language understanding and imagine that a more refined language model can help here. So far, no large language model is publicly available that includes EHR-style text in its training dataset.

Classic Ranking & Recommendation Approaches Fail. Designing and implementing a CDSS process is a complex task. Therefore, we choose well-known classical methods for cohort selection and recommendation ranking. Our results show that these methods sometimes deliver unsatisfying results.

Overall. We conclude that the deep learning-enabled DDx process is generally applicable and leads to an understandable and transparent system that helps clinicians obtain a better situation understanding. Medical practitioners want to work with the system in the future and report that the system made helpful recommendations and enabled better situation awareness. Thus, we conclude that HYPOTHESIS H3 is confirmed. In essence, we validate the proposed DDx process.

5.2.5.2. Design Challenges for Text-based CDSS

We identify the following challenges for designing and implementing a text-based DDx process based on our observations.

System Success is Bound to Understandable User Experience Design. User experience design, medical domain expertise, natural language process, and deep learning proficiency are the primary building blocks for a successful text-based CDSS. Consequently, even research prototypes need to deliver good results for all

three of these building blocks to obtain valuable evaluation data. Failing at one of these key components can lead to unusable observations and results.

Users Require a Transparent and Comprehensible Recommendation Process to Trust. Understanding the reasoning of a CDSS is essential to many clinicians to build trust in the process and system. Fulfilling this requirement becomes more challenging with model complexity. Our experiment outlined the need for better models, such as deep learning-based approaches. Deep learning models' reasoning is hard to interpret, which characterizes a future research opportunity.

Lack of Large-scale Clinical Language Understanding Models for Research. Many language understanding tasks benefit from large language models, such as GPT-3 (Brown et al., 2020) or BERT (Devlin et al., 2019). Gu et al., 2021 have shown that language models trained on in-domain data work better for biomedical topics than general-purpose models. However, the publicly available biomedical language models (Gu et al., 2021; J. Lee, Yoon, et al., 2019; Peng, S. Yan, and Lu, 2019) are trained on research papers taken from PubMed. Therefore they are still out-of-domain for clinical texts.

Lack of Universal Applicable Time-aware Patient Representations for Research. Language understanding alone is not enough to obtain a holistic patient representation. EHRs consist of time-bound medical events and actions (Glicksberg, Miotto, et al., 2018; Miotto, L. Li, et al., 2016). This property is not accounted for when using a language model alone for patient representation. Landi et al., 2020; Miotto, L. Li, et al., 2016; Miotto, F. Wang, et al., 2018 proposed models to counteract this shortcoming, but still, those models are not freely available for researchers.

Incorporating Medical Knowledge from Multi-modal Sources is Unsolved. For a comprehensive representation of a patient situation, it is necessary to incorporate information from clinical language understanding models working on EHRs, time-aware patient representation models, and knowledge from external sources such as knowledge bases (UMLS), textbooks, and research papers. Tackling this challenge will increase the soundness of recommended clinical action

vastly. However, it is still unclear how this diverse information set can be effectively combined and used in deep learning-based recommendation models.

No Large-scale Datasets are Available for Research. One of the most significant challenges is the lack of large-scale data. The amount of datasets available for research in CDSS is minimal. The available datasets are rather small and focus on medical subfields such as intensive care (Johnson et al., 2016) or oncology (Borchert et al., 2020).

No Standardized Evaluation Framework for CDSS and DDx Systems. Defining standards for language understanding capabilities, the resulting tasks, and their evaluation have considerably amplified research on language understanding (A. Wang, Pruksachatkun, et al., 2019; A. Wang, Singh, et al., 2018). For CDSS, there exist no standard processes, a definition of capabilities, or commonly occurring tasks. Lately, there has been some work on creating biomedical text understanding benchmarks (Gu et al., 2021; Peng, S. Yan, and Lu, 2019), but those lack a focus on clinically relevant applications. We encourage the community to develop an openly available, standardized evaluation framework for text-based CDSS to enable comparability and more rapid advancement of the field.

5.2.5.3. Limitations

Given the setup of our study, we report the following limitations. We conducted our experiments using the relatively small ICU dataset MIMIC-III as a single source of knowledge. The complete hospital archive would be available to the SmartMD system in an ideal real-world application. Moreover, medical experts would perform diagnostics to refine their understanding of the patient's situation.

Clinicians often have busy schedules. Therefore, we were only able to recruit five participants for our experiments. We found that the participants disagreed in some aspects, limiting our findings' generality. As a result, our qualitative investigations pose valuable groundwork for more extensive empirical studies and to further improve the Deep Learning-aided DDx process.

5.2.6. Background

Computer-aided CDSS have been in the focus of research for multiple decades until now. R. T. Sutton et al., 2020 performed a meta-analysis of their benefits, risks, and success strategies from 1980 until 2018. Thereby they categorize the approaches by scope, for example, Patient Safety (Eslami et al., 2012; Mahoney et al., 2007; McEvoy et al., 2017), Clinical Management (McMullin et al., 2004 Sep-Oct; Salem et al., 2018), Diagnostic Support (Cui, Bozorgi, et al., 2012; De Fauw et al., 2018; Ronicke et al., 2019) or Patient Decision Support (Jungmann et al., 2019) and method into rule-based (Cui, Bozorgi, et al., 2012) or trained models (Arandjelović, 2015; Bakator and Radosav, 2018; Miotto, F. Wang, et al., 2018). Much related work focuses on a specific disease (De Fauw et al., 2018; Goldenberg, Nir, and Salcudean, 2019; D. Jiang et al., 2020; Y. Liu et al., 2019) while we aim for a generally applicable system that uses trained models.

5.2.6.1. Text-based Clinical Decision Support Systems

A branch of Clinical Decision Support systems exploits text data stored in EHRs (Shickel et al., 2018). Common tasks that benefit from this data are clinical outcome prediction (J. Lee, Maslove, and Dubin, 2015; van Aken, Papaioannou, et al., 2021), medical coding assistance (Bell, Jalali, and Mensah, 2013; Catling, Spithourakis, and Riedel, 2018; Shi, 2017) or discovering disease subtypes (L. Li et al., 2015). The task of clinical reasoning, especially the differential diagnosis process, can benefit from the analysis of EHRs. Differential diagnosis involves critical exploration of patient history, physical examination, and careful review of the data obtained in laboratories and diagnostic image settings (Altkorn, 2020; Croskerry, 2009).

5.2.6.2. Clinical Natural Language Understanding

For text-based CDSS, such as SmartMD, is language understanding an important prerequisite and often consists of multiple pre-processing steps, such as clinical concept recognition, clinical concept linking, and negation detection.

Clinical Concept Recognition & Linking. Clinical Concept Recognition is the task of identifying words that mention a clinical concept in a text (Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017; Si et al., 2019; Y. Wu, M. Jiang, et al., 2018). Assigning those mentions to concept identifiers in standardized ontologies, such as UMLS (Bodenreider, 2004), is the task of clinical concept linking (Aronson and Lang, 2010 May-Jun; Fu et al., 2020). Both concept recognition and linking models have to deal with a broad set of challenges arising from the clinical setting. For example, clinical notes in EHRs are often taken with pressing time constraints and come with special challenges such as flexible formatting, atypical grammar, misspellings, and ad-hoc abbreviations, which also tend to be specific to hospital departments (Leaman, Khare, and Lu, 2015). As a consequence, the literature proposes a wide variety of methods comprising linguistically motivated, rule-based approaches (Aronson and Lang, 2010 May-Jun; Tseytlin et al., 2016), compositional mixture-of-experts-like (Yuksel, J. N. Wilson, and Gader, 2012) approaches (D'Souza and V. Ng, 2015; Rajani, Bornea, and Barker, 2017), as well as shallow (Leaman and Lu, 2016) and deep learning (Choi, Chiu, and Sontag, 2016; Jauregi Unanue, Zare Borzeshi, and Piccardi, 2017; Mueller and Durrett, 2018; Schumacher and Dredze, 2019; Schumacher, Mulyar, and Dredze, 2020; Y. Wu, M. Jiang, et al., 2018) models (Fu et al., 2020). In this work, we use an LSTM-based concept recognition system proposed by (Arnold, Gers, et al., 2016).

Negation Detection. The meaning of clinical concepts is heavily affected by modifiers such as negation or uncertainty (Mehrabi et al., 2015; Uzuner et al., 2011). Accordingly, is negation detection a complex task that requires a vast understanding of the context in which a clinical concept appears to extract actionable knowledge from clinical text (Chapman, Bridewell, et al., 2001; Cotik et al., 2016). One of the earliest approaches to Assertion Detection is NegEx (Chapman, Bridewell, et al., 2001), which used handcrafted extraction patterns to recognize the *absent* class. Recently, neural network architectures have been applied to solve this problem. For example, Qian et al., 2016 approach this challenge Convolutional Neural Networks (Lecun et al., 1998), while Sergeeva et al., 2019 propose a model based on Long-Short Term Memory (Hochreiter and J. Schmidhuber, 1997) with forget gates (Gers, J. A. Schmidhuber, and Cummins, 2000). Most recently, van Aken, Trajanovska, et al., 2021 and Alsentzer et al., 2019 explored also the

capabilities of transformer-based (Vaswani et al., 2017) architectures.

Cohort Modeling. Modeling and retrieving cohorts is commonly performed in clinical research to find a group of patients that share a set of discriminative features. Consequently, the literature proposes a wide variety of patient characteristics, such as symptoms, comorbidities, demographics, and treatments, to represent the patients' condition for particular applications (Cui, Bozorgi, et al., 2012; L. Li et al., 2015; Sharafoddini, Dubin, and J. Lee, 2017). Aiming to automate the cohort selection process, approaches reported in the literature date back to 1989, as reported by Sharafoddini, Dubin, and J. Lee, 2017. In addition to clinical research, using cohorts of similar patients to recommend personalized treatments, diagnoses or diagnostics has become an emerging topic in recent research (Sharafoddini, Dubin, and J. Lee, 2017). For example, Cui, Bozorgi, et al., 2012 construct cohorts for epilepsy research based on clinical concepts extracted from EHRs and hand-crafted extraction rules. J. Lee, Maslove, and Dubin, 2015 show that outcome prediction models perform better when trained on relevant cohorts. To do so, they calculate a cosine-distance-based measure on manually crafted feature vectors. Glicksberg, Miotto, et al., 2018 propose a semi-supervised approach to compute a word2vec-based (Mikolov, K. Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) clinical concept embedding to group patients into cohorts. Following this idea, Miotto, L. Li, et al., 2016 and Landi et al., 2020 propose a scalable unsupervised patient representation based on autoencoders (Kramer, 1991) and convolutional neural networks (Lecun et al., 1998) that enables the clustering of patients into cohorts by latent clinical features.

5.3. Conclusion

This chapter addressed RQ4: "How Effective are Deep Learning Enhanced Medical Information Seeking Processes?" by investigating two typical medical processes enhanced with deep learning models. First, we presented neural entity recognition and topic classification models for medical passage retrieval (SmartMD: IR). Second, we designed the deep learning-enabled differential diagnosis process (SmartMD: DDx).

We demonstrated that neural topic models combined with entity recognition

models enable clinicians to search for topical facets of diseases. Using deep-learning-based models to retrieve the most relevant paragraphs for clinicians can reduce the time needed for research by selecting coherent passages. Furthermore, neural topic models such as SECTOR enable clinical decision support systems to select answer passages based on semantic similarity without being restricted to lexical features.

We presented "SmartMD: DDx," a text-based clinical decision support system. We designed a deep learning-aided differential diagnosis process implemented in the SmartMD system. We utilize deep learning models for text understanding, classical methods for clinical action recommendation, and cohort selection techniques that incorporate clinicians' input.

Our study with five medical professionals validates the deep-learning-enabled differential diagnosis process. The participants worked on two randomly drawn cases from the MIMIC-III dataset using the SmartMD: DDx system. Our experiments revealed seven significant design challenges for clinical decision support systems on this foundation. Especially the need for a benchmark on a standardized set of capabilities focused on CDSS arose as a valuable asset for future research. Our results show that deep learning models can benefit text-based clinical decision support systems. Particularly abstracting, operationalizing, and standardizing well-known clinical processes can result in beneficial outcomes, as our deep-learning-enabled differential diagnosis process shows.

Chapter 6

Conclusion and Future Work

This thesis analyzed the suitability of discrete and neural text representations in text-based clinical decision support systems. We have designed the deep learning-enabled differential diagnosis process as an exemplary clinical application of machine reading models. Accordingly, we contributed and evaluated the SmartMD: IR and SmartMD: DDx systems that implement this process.

Based on this deep learning-enabled differential diagnosis process, we derived requirements for neural text representation to be suitable for application in CDSS. During the design process, we have explored the properties of Open Information Extraction systems. Therefore, we created the RelVis benchmark and used it to show the shortcomings of Open Information Extraction systems that interfere with their application in CDSS. As an alternative, we proposed and investigated neural text representations. Alongside, we created in Arnold, Schneider, et al., 2019 SECTOR, a neural topic model that can extract coherent passages from long documents in a medical literature search.

The deep learning-enabled differential diagnosis process requires text representations to capture a holistic understanding of clinical documents, such as EHRs and medical literature. Fulfilling this requirement mandates focusing on multiple textual modalities, such as topic, local context, global context, and medical concept resolution. Therefore, we evaluated the compositionality of neural text embeddings and found that differing pretraining goals lead to complementary text representations.

This chapter reviews our contributions regarding the desired properties of the deep learning-enabled differential diagnosis process (Section 6.1). In Section 6.2, we review the research questions and our answers. Next, we describe our prior assumptions and limitations of our work in Section 6.3. Section 6.4 summarizes our work from the perspective of business opportunities. Finally, we discuss research questions revealed by our work in Section 6.5.

6.1. Contributions

We introduced central challenges for text-based clinical decision support systems in Section 1.4. We designed the Deep Learning enabled Differential Diagnosis Process to provide an overarching application scenario with those challenges in mind. Accordingly, we designed and evaluated possible solutions for subtasks in this framework. In the following, we discuss our contributions and findings:

Deep Learning enabled Differential Diagnosis Process. We approach the broad field of text-based clinical decision support systems by focusing on Differential Diagnosis support. Therefore, we design the Deep Learning enabled Differential Diagnosis Process by abstracting the clinical process. (Section 5.2.2) All six steps of this process formalization can be solved using statistical and machine learning models. We design and implement SmartMD: IR (Section 5.1), a medical passage retrieval system, and SmartMD: DDx (Section 5.2), a differential diagnosis support system, accordingly.

In-depth Analysis of Open Information Extraction Systems. Clinical use cases require high scalability in runtime, and it is also crucial to maintain high precision and recall. An inherent property of the Open Information Extraction paradigm is its scalability to large amounts of text data. We have advanced this property by integrating the OIE paradigm in a fast in-memory database system and reporting execution times in seconds. (Section 3.1) Our system can efficiently integrate pre-existing knowledge held in a database with insights from text. (Section 3.1.4)

Our in-depth analysis of OIE Systems (Section 3.2) revealed a lack of stringent annotation policies, making a comparative analysis and design of OIE systems challenging. Moreover, we have observed that each tested OIE system depends

on syntactic taggers that often propagate errors down to the logic for extracting OIE tuples. The benchmarked systems often extract unnormalized relation tuples that do not leverage the well-researched concept of "normal forms" in database theory. We have further found that the reviewed OIE systems, which are already overfitting and struggling with news datasets, will likely have issues achieving the needs for application in idiosyncratic domains such as clinical narratives. (Section 3.2.6)

Open Information Extraction Benchmark. We have created RelVis a benchmark suite to overcome the challenges in evaluating the quality of OIE Systems. (Section 3.2.5) RelVis was the first benchmark that combined four labeled datasets and supported the five most recent[1] OIE systems. RelVis allows performing both qualitative and quantitative analyses.

Coherent Medical Topic Segmentation. We have introduced the new "PubmedSection," dataset (Section 4.3) for evaluating coherent topic segmentation models on medical texts. Accordingly, we have adapted the SECTOR model to capture medical information using the PubmedSection dataset. (Section 4.4)

Holistic Text Representations for Medical Applications. We have identified effective combinations of universal and specialized text embeddings in an extensive study on 11 tasks. (Section 4.1) We have extended SentEval to the medical domain by integrating the "WikiSection.diseases" and the "PubMedSection" task. (Section 4.3) Our comprehensive analysis reports that combining universal and specialized embeddings, such as ELMo + SECTOR, yields improved results in many downstream tasks. (Section 4.4) Furthermore, combining complementary embedding combinations yield holistic text representations that achieve a new state-of-the-art for two tasks in SentEval. (Section 4.4.2)

Medical Literature Search System. We have demonstrated that neural topic models combined with entity recognition models enable clinicians to search for topical facets of diseases. Furthermore, we report that neural topic models such as

[1] At the time of writing of (Schneider, Oberhauser, Klatt, et al., 2017a; Schneider, Oberhauser, Klatt, et al., 2017b)

SECTOR enable clinical decision support systems to select answer passages based on textual semantic similarity without being restricted to lexical features. (Section 5.1)

6.2. Review of Research Questions

RQ1: Is the Open Information Extraction Paradigm Suitable for Clinical Text Understanding? Clinical decision support systems require text understanding systems to deliver representations of clinical concepts with high recall, especially when encountering rare concepts. Moreover, these representations must handle syntactic and semantic errors in texts while being adaptable and scaleable to new concepts and large volumes of documents. First, we identified the scalability of the Open Information Extraction paradigm as a central property to be successfully applied in clinical settings. We designed INDREX-MM, an Open Information Extraction system that combines the scalability of in-memory database systems with the high-recall focused relation extraction capabilities of OIE systems. We have shown that INDREX-MM performs relation extraction within seconds of execution time on 800.000 documents. Another essential advantage of INDREX-MM is straightforward integration with existing knowledge bases using SQL statements. Secondly, we tackled the lack of integrated benchmarks for OIE systems, posing a significant challenge when comparing OIE systems. We designed RelVis, the first benchmarking system for OIE that incorporates four well-known datasets and supports four OIE systems. RelVis supports exact and weak match strategies on the annotation level to compare OIE systems with differing annotation styles. Our system allows quantitative automated and manual qualitative evaluations and supports human judges by classifying errors.

Third, we evaluated four OIE systems using RelVis. We discovered that OIE systems have diverging annotation styles. Thus, there is disagreement within the research community on a clear task definition. Moreover, we revealed that the surveyed systems tend to overfit specific data sets within the general news domain. None of the evaluated systems uses the well-researched "normal forms" from database theory which often leads to ambiguous extraction, especially for n-ary relations. Most surveyed systems depend on linguistic features obtained in a preprocessing stage. Errors in this stage propagate downstream, which is a

considerable disadvantage in clinical documents that often contain writing errors. Following our experiments, we conclude that the surveyed OIE systems must overcome these issues to be suitable for clinical applications.

RQ2: Can Neural Text Representations aid Text-based Clinical Decision Support Systems? An essential capability of clinical text representations is to capture the global context describing a patient's situation. A clinical text representation must be able to identify coherent passages in medical texts, recognize their topical facet, and capture a single sentence's meaning given the context of the entire document. We designed in Arnold, Schneider, et al., 2019 the coherent medical topic segmentation task to address this challenge with a measurable target. We designed the novel "PubmedSection" dataset accordingly and benchmarked the SECTOR model. Overall, we conclude that the SECTOR model is a robust and extensible building block representing medical texts.

RQ3: Are Text Representations Trained with Differing Pretraining Goals Complementary? Identifying and combining complementary text representations can lead to holistic representations that improve medical language understanding systems. We investigated if and how text representations with differing pretraining, such as Language Modeling, Topic Modeling, or Entity Linking tasks, differ in their capabilities. We identified complementary pairings of general-purpose and specialized text representations. Our comprehensive analysis resulted that combining universal and specialized embeddings, such as ELMo + SECTOR, yields considerably improved results in many downstream tasks. Furthermore, we showed that complementary combinations yield holistic text representations that achieve a new state-of-the-art for two tasks in SentEval.

RQ4: How Effective Are Deep Learning Enhanced Medical Information Seeking Processes? Clinicians need to access state-of-the-art medical literature and medical experience gained over the years. These time-consuming tasks benefit from clinical decision support systems, which incorporate the experiences of medical personnel in the clinic collected in EHRs and the most recent literature.

We presented SmartMD:IR, a medical passage retrieval system. SmartMD:IR combines neural topic models with entity recognition models and enables clinicians

to search for topical facets of diseases. We designed a deep learning-aided differential diagnosis process implemented in the SmartMD: DDx system. SmartMD: DDx utilizes deep learning models for text understanding and classical clinical action recommendation and cohort selection methods that incorporate clinicians' input.

Our two-fold experiment with five medical professionals validates the deep-learning-enabled differential diagnosis process and evaluates our proof of concept implementation. Our user observation study and a structured interview validate the effectiveness of the deep-learning-enabled differential diagnosis process. Moreover, we have revealed seven significant design challenges for clinical decision support systems. Subsequently, we conclude that text-based clinical decision support systems can benefit from deep learning models.

6.3. Limitations

We have presented clinical text-understanding methods and deep learning-aided text-based differential diagnosis support systems. Driven by confining factors and resources, our work focuses on applicability in European countries. Therefore, our approach is subject to limitations that the future work could approach.

Limited Evaluation Datasets. We evaluated our methods on a manifold of datasets, some out of the medical domain. This evaluation regime allowed us to prove the transferability of our developed models, circumventing the lack of publically available clinical datasets. Our evaluation of the medical domain often used medical literature as a surrogate for clinical texts. Unfortunately, we could only evaluate our methods using one clinical dataset, MIMIC-III. Due to strict data protection laws, MIMIC-III is the only available research dataset featuring complete EHRs. Therefore, we could not validate the transferability of our methods to other clinics, medical subjects, or languages that MIMIC-III does not include.

Digital Text in EHRs Required. Our methods assume that health records are digitally available. Many hospitals have implemented EHRs by now, but non-digital data can not be handled by our methods without prior digitalization steps, such as optical character recognition.

Focusing on Text-Representations. We aimed to create medical text representations that capture multiple textual modalities as a building block for a patient representation. However, a holistic patient representation demands time-aware multi-modal representations that incorporate text, structured data, and medical imaging. Accordingly, are the models proposed in this thesis a building block in the direction of a holistic patient representation.

Model Drift and Continuous Learning. Clinical decision support systems need to address the newest insights on diseases, diagnostics, treatments, and therapies. Therefore, model drift has to be expected over time and compensated. In our work, we described methods to train models initially. However, it can be beneficial to train models in an iterative human-in-the-loop manner based on user feedback. We have not focussed on this topic and leave it open for future work.

6.4. Business Perspectives

Deep Learning-based clinical text understanding methods open up a broad range of Business Opportunities. Due to the text-based nature of medical documentation, note-taking, and research publication, it is imperative to develop systems that help medical professionals manage the growing amount of medical texts. Methods developed for Text-based clinical decision support systems apply to a wide range of medical business cases. Indeed, the broad adoption of EHRs and the collection of structured measurements together with unstructured clinical narratives and medical imaging was a big step forward to improving treatment quality and optimizing processes. However, important information collected in clinical narratives is still unused. For example, medical professionals could provide more effective and efficient care if a clinical support system provides insights based on experience from similar cases documented in a hospital's archive. Moreover, connecting archival data with clinical process coordination could improve clinical pathways, enrich patient triage, and increase revenue via clinical coding assistance systems.

6.4.1. Clinical Action Recommendation

Identifying a Patient's situation and trajectory is a central task for medical professionals. This complex process requires medical professionals to pay attention to many textual sources, including doctors' letters, existing medical and nursing notes, and medical literature. Connecting knowledge stored in textual data to clinical decision-making can improve treatment quality and the efficiency and effectiveness of clinical pathways. The central point is to provide medical practitioners with a context-aware recommendation regarding which literature to read and action to take given a patient's situation, predicted trajectory, and cohort affiliation. Neural text representations are an essential building block to make information stored in EHRs accessible for further processing. CDSS can use the resulting enriched EHRs to perform context-sensitive literature searches or recommend clinical pathways.

The required neural text representation models can obtain general language understanding knowledge on public data and transfer it to the clinical domain by finetuning the model on smaller datasets of clinical archives. The resulting models enable CDSS to search and cluster medical narratives and build a strong foundation for downstream classification and recommendation models. Clinical neural text representations open the opportunity to connect unused clinical data directly with medical decision-making and clinical pathway optimization.

6.4.2. Clinical Coding

Besides supporting medical doctors and caretakers, clinical text understanding models can also assist clinical coders in their day-to-day work. Clinical coding involves understanding clinical narratives to issue correct bills for medical care compensation. This task requires highly trained professionals to read and understand many multi-document EHRs to select the correct billing codes. Accordingly, this task is time-consuming and error-prone. A neural clinical text representation model could assist in predicting billable clinical codes. These can be applied automatically for straightforward situations and direct medical coders' attention to complex cases. In such situations, a deep learning-based model can support clinical coders by gathering details that might span multiple documents. As a result, such models can reduce errors and time needed per document and increase the recall of billable codes. Accordingly, coding assistance systems can increase

revenue while increasing efficiency per clinical coder.

6.5. Future Work

Our research on deep learning-based clinical language understanding for differential diagnosis support revealed questions and perspectives for future research. Further, we identified research problems that we considered out of focus for this Thesis. We discuss in this section the three most important of those questions as a possible direction for future research.

Holistic Combination of Text Representations with Differing Modality

Universal text representations, such as ELMo (M. E. Peters et al., 2018), BERT (Devlin et al., 2019), or GPT (Radford, J. Wu, Amodei, et al., 2019; Radford, J. Wu, Child, et al., 2019), rely for training on variations of the neural probabilistic language model (Bengio et al., 2003). Specialized embeddings, such as presented by Arnold, Schneider, et al., 2019; Arnold, van Aken, et al., 2020, Pappu et al., 2017, M. Chen et al., 2019 and Landi et al., 2020 use specialized pretraining goals to obtain text representations focused on selected properties. Our work has shown that universal text representations miss details expressed in the encoded text that specialized representations can capture.

From this observation arises the desire to understand how these representations differ and the root cause of the differences. Understanding these aspects might improve text representations vastly. Naturally, the follow-up is how we can obtain holistic text representations. Naive approaches to this problem do not deliver satisfying results. Concatenating text representation vectors leads to an increase in dimensions and, therefore, additional computational complexity. Coates and Bollegala, 2018 propose averaging source representations that work surprisingly well for a few representations but lose information at scale. We observe three lines of research in this direction. First, Pfeiffer et al., 2021 approach this problem by augmenting large language models using adapters as plug-in components. Secondly, intrinsic learning by multi-task pretraining (Raffel et al., 2020). Third, extrinsic learning by meta-learning (Bollegala, Hayashi, and Kawarabayashi, 2018; Kiela, C. Wang, and Cho, 2018; Rettig, Audiffren, and Cudré-Mauroux, 2019;

L. Wang, Y. Li, and Lazebnik, 2016).

Time-Aware Multi-Modal Patient Representations

Incorporating textual information from EHRs, time-aware patient representation models, and knowledge from external sources such as knowledge bases (UMLS), textbooks, and research papers is crucial to obtain a text-based patient representation. Language understanding alone is not enough to obtain a holistic patient representation. EHRs consist of time-bound medical events and actions. This property is not accounted for when using a text representation alone. Landi et al., 2020; Miotto, L. Li, et al., 2016; Miotto, F. Wang, et al., 2018 proposed models to counteract this shortcoming. However, this representation still does not depict imaging results, medical signals, and structured data. There is research on each modality independently (Miotto, F. Wang, et al., 2018), but we still lack a multi-modal, time-aware patient representation. Moreover, it is unclear how to combine these independent representations efficiently to serve as a unified foundation for clinical decision-support models.

Interdisciplinary Evaluation of the Deep Learning Aided DDx Process

Generating value with data products is a complex and demanding endeavor. Verifying the effectiveness of deep learning-aided DDx systems requires similar steps. It is crucial to evaluate DDx systems in end-to-end tests in a clinical study to verify their effectiveness. Bernardi, Mavridis, and Estevez, 2019 have shown that model quality does not linearly translate into increased KPIs. Moreover, we have seen in our evaluation that UX design influences the perceived model and process performance to a large extent which was also described by Smith et al., 2018. As a result, we raise the need for an interdisciplinary end-to-end evaluation of the Deep Learning aided DDx process involving researchers from the fields of user experience design, deep learning, and medical research.

Opportunities for Large Language Models in CDSS

Incorporating large language models (LLMs) such as PaLM 2 (Anil et al., 2023) or med-palm (Singhal et al., 2023) into the healthcare system can unlock gains in

efficiency and patient engagement. LLM-based tools are proficient at interpreting and generating natural language, opening up many opportunities to enhance the delivery of medical care.

In medical research, LLM-based chatbots can be used to condense the latest medical research. Especially in combination with retrieval augmented generation methods (Borgeaud et al., 2022), these systems can distill complex scientific papers into summaries, recommend papers to read, and answer questions. These systems can reduce the time healthcare professionals require to access the latest research.

Training LLMs to anonymize sensitive data and to replace personal identifiers in medical documents with pseudonyms to protect patient confidentiality can enable data sharing between actors in the healthcare system at a large scale. As a result, researchers can access more relevant data to design the next generation of CDSS.

Patients can benefit from specialized LLM models that break down the barriers to information by translating medical "doctor language" into straightforward terms, allowing patients to gain more detailed insights into their health conditions. Furthermore, LLM-based chatbots can operate as symptom-collection systems. Those systems could employ a methodical approach to asking patients for their medical history in natural language. This comprehensive, streamlined approach to data collection can augment the patient's health profile, ensuring clinicians have a detailed understanding upon which to base their diagnosis and treatment plans.

Another promising application of LLMs lies in their ability to efficiently summarize and extract pertinent information that is spread across multiple EHR entries and documents. Allergies, past medical histories, and current medications can be quickly identified, ensuring that doctors have immediate access to crucial health information. Additionally, LLM writing assistants have the potential to code clinical narratives on the fly into standardized medical formats like ICD or FHIR, promoting seamless integration and improving interoperability across the healthcare landscape.

Acknowledgements

Completing this thesis would not have been possible without the support and guidance of many individuals.

First and foremost, I would like to express my sincerest gratitude to my supervisor, Alexander Löser, for his tireless support and guidance throughout this journey. From the very beginning of my research at the Berliner Hochschule für Technik, Alexander has been a constant source of knowledge, expertise, and encouragement. The ice cream we had during our first discussion marks the beginning of a journey that has led to this thesis, and I am deeply thankful for Alexander's invaluable insights and guidance in the fields of natural language processing and deep learning. His constant encouragement and support kept me motivated through even the most challenging times.

I would also like to thank the members of my thesis committee, Steffen Staab and Georg Rehm, for their invaluable feedback and suggestions on the various drafts of this thesis. Their expertise and perspectives have greatly enhanced my research and provided valuable guidance. I sincerely appreciate their support and am grateful for the opportunity to work with such esteemed scholars.

I want to express particular gratitude to Felix A. Gers, Amy Siu, and Peter Tröger, without whom the research group would lack critical parts, whether providing the technical infrastructure at DATEXIS or crucial feedback on my research ideas during this turbulent journey.

I am grateful to my colleagues at DATEXIS for providing a supportive and collaborative environment. In particular, I would like to thank my co-authors Sebastian Arnold, Klemens Budde, Philippe Cudré-Mauroux, Jens Graupmann,

Paul Grundmann, Cordula Guder, Torsten Kilias, Tobias Klatt, Oleksandr Kozachuk, Manuel Mayrdorfer, Tom Oberhauser, Henriette Schmidt, and Thomas Steffek for their help, discussions, and advice on various aspects of my research. Likewise, I want to thank Robert Dziuba, Alexei Figueroarosero, Phillip Haustein, Jens-Michalis Papaioannou, Katharina Sachs, Janine Schleicher, Betty van Aken, and Benjamin Winter. Their support, feedback, and critical discussions have been crucial in navigating the challenges that came with this journey.

Furthermore, I would like to thank my family and friends who accompanied me along this way and made this journey possible. I'm incredibly thankful for the aid of Rike Daetz, Laura Hartgers, Pia Kasper, Henryk and Maria Plötz, and Rebecca Start. Their belief in me has been a constant source of strength, and I am grateful to have such a wonderful network of people in my life.

I dedicate this thesis to my parents, Bernd and Martina Schneider, and my grandparents, Dieter and Rosemarie Dressler, for their support and encouragement throughout my academic journey. They have been my rock, and I am grateful for their love and guidance.

Bibliography

Akbik, A., D. Blythe, R. Vollgraf (Aug. 2018). 'Contextual String Embeddings for Sequence Labeling'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649 (cit. on p. 63).

Akbik, A., J. Bross (Apr. 2009). 'Wanderlust: Extracting Semantic Relations from Natural Language Text Using Dependency Grammar Patterns'. In: *Proceedings of the 2009 Semantic Search Workshop at the 18th International World Wide Web Conference*. SemSearch '09. Madrid, Spain, pp. 6–15 (cit. on p. 25).

Akbik, A., A. Löser (2012). 'Kraken: N-ary Facts in Open Information Extraction'. In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*. Association for Computational Linguistics, pp. 52–56. (Visited on 06/23/2016) (cit. on pp. 24, 25, 43).

Alsentzer, E., J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott (June 2019). 'Publicly Available Clinical BERT Embeddings'. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 72–78. arXiv: 1904.03323 (cit. on pp. 107, 131).

Altkorn, D. (2020). 'A Model for Clinical Reasoning'. In: *Symptom to Diagnosis: An Evidence-Based Guide*. Ed. by S. D. Stern, A. S. Cifu, D. Altkorn. 4th ed. New York, NY: McGraw-Hill Education. (Visited on 08/28/2022) (cit. on pp. 21, 100, 101, 130).

Angeli, G., M. J. Premkumar, C. D. Manning (2015). 'Leveraging Linguistic Structure for Open Domain Information Extraction'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 344–354. (Visited on 07/18/2016) (cit. on pp. 25, 41, 47, 51).

Anil, R. et al. (Sept. 2023). *PaLM 2 Technical Report*. arXiv: 2305.10403 [cs]. (Visited on 11/05/2023) (cit. on p. 144).

Arandjelović, O. (Dec. 2015). 'Discovering Hospital Admission Patterns Using Models Learnt from Electronic Hospital Records'. In: *Bioinformatics* 31.24, pp. 3970–3976. (Visited on 04/14/2020) (cit. on pp. 19, 110, 111, 130).

Arnold, S., R. Dziuba, A. Löser (2016). 'TASTY: Interactive Entity Linking As-You-Type'. In: *COLING'16 Demos*, pp. 111–115. (Visited on 01/18/2017) (cit. on pp. 88, 94, 105).

Arnold, S., F. A. Gers, T. Kilias, A. Löser (2016). 'Robust Named Entity Recognition in Idiosyncratic Domains'. In: *arXiv:1608.06757 [cs.CL]*. arXiv: 1608.06757 [cs.CL] (cit. on pp. 88, 94, 99, 104, 131).

Arnold, S., R. Schneider, P. Cudré-Mauroux, F. A. Gers, A. Löser (Mar. 2019). 'SECTOR: A Neural Model for Coherent Topic Segmentation and Classification'. In: *Transactions of the Association for Computational Linguistics* 7, pp. 169–184. (Visited on 06/05/2019) (cit. on pp. 12, 14, 63, 65–67, 70, 74, 76, 78, 87, 99, 135, 139, 143).

Arnold, S., B. van Aken, P. Grundmann, F. A. Gers, A. Löser (Apr. 2020). 'Learning Contextualized Document Representations for Healthcare Answer Retrieval'. In: *Proceedings of The Web Conference 2020*. WWW '20. New York, NY, USA: Association for Computing Machinery, pp. 1332–1343. arXiv: 2002.00835. (Visited on 12/09/2022) (cit. on p. 143).

Aronson, A. R., F.-M. Lang (2010 May-Jun). 'An Overview of MetaMap: Historical Perspective and Recent Advances'. In: *Journal of the American Medical Informatics Association: JAMIA* 17.3, pp. 229–236 (cit. on pp. 6, 23, 131).

Arora, S., Y. Liang, T. Ma (2017). 'A Simple But Tough-To-Beat Baseline for Sentence Embeddings'. In: *ICLR*. (Visited on 06/30/2017) (cit. on p. 73).

Aspland, E., D. Gartner, P. Harper (Jan. 2021). 'Clinical Pathway Modelling: A Literature Review'. In: *Health Systems* 10.1, pp. 1–23. (Visited on 08/26/2021) (cit. on pp. 2, 19).

Bakator, M., D. Radosav (2018). 'Deep Learning and Medical Diagnosis: A Review of Literature'. In: *Multimodal Technologies and Interaction* 2.3, p. 47 (cit. on pp. 19, 130).

Balaneshin-kordan, S., A. Kotov (2018). 'Deep Neural Architecture for Multi-Modal Retrieval Based on Joint Embedding Space for Text and Images'. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. New York, NY, USA: ACM, pp. 28–36. (Visited on 06/12/2018) (cit. on p. 34).

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni (2007). 'Open Information Extraction from the Web.' In: *IJCAI*. Vol. 7, pp. 2670–2676. (Visited on 07/05/2016) (cit. on pp. 4, 7, 24, 49, 50).

Beeferman, D., A. Berger, J. Lafferty (1999). 'Statistical Models for Text Segmentation'. In: *Machine learning* 34.1, pp. 177–210. (Visited on 05/03/2017) (cit. on p. 6).

Bell, C. M., A. Jalali, E. Mensah (2013). 'A Decision Support Tool for Using an ICD-10 Anatomographer to Address Admission Coding Inaccuracies: A Commentary'. In: *Online Journal of Public Health Informatics* 5.2, p. 222 (cit. on pp. 21, 130).

Bellazzi, R., F. Ferrazzi, L. Sacchi (Sept. 2011). 'Predictive Data Mining in Clinical Medicine: A Focus on Selected Methods and Applications'. In: *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery* 1, pp. 416–430 (cit. on p. 19).

Bengio, Y., R. Ducharme, P. Vincent, C. Janvin (2003). 'A Neural Probabilistic Language Model'. In: *Journal of Machine Learning Research* 3, pp. 1137–1155 (cit. on pp. 29, 30, 143).

Bernardi, L., T. Mavridis, P. Estevez (2019). '150 Successful Machine Learning Models: 6 Lessons Learned at Booking.Com'. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, pp. 1743–1751 (cit. on p. 144).

Berner, E. S., ed. (2007). *Clinical Decision Support Systems: Theory and Practice*. 2nd ed. Health Informatics. New York, NY: Springer (cit. on pp. 64, 67).

Bhatia, S., J. H. Lau, T. Baldwin (2016). 'Automatic Labelling of Topics with Neural Embeddings'. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 953–963 (cit. on p. 33).

Blei, D. M. (Apr. 2012). 'Probabilistic Topic Models'. In: *Communications of the ACM* 55.4, p. 77. (Visited on 05/22/2019) (cit. on pp. 4, 6, 33, 68).

Blei, D. M., A. Y. Ng, M. I. Jordan (2003). 'Latent Dirichlet Allocation'. In: *Journal of Machine Learning Research* 3.Jan, pp. 993–1022. (Visited on 02/23/2017) (cit. on pp. 33, 91).

Boden, C., A. Löser, C. Nagel, S. Pieper (2011). 'FactCrawl: A Fact Retrieval Framework for Full-Text Indices.' In: *WebDB*. (Visited on 12/08/2016) (cit. on p. 24).

Bodenreider, O. (Jan. 2004). 'The Unified Medical Language System (UMLS): Integrating Biomedical Terminology'. In: *Nucleic Acids Research* 32.Database issue, pp. D267–D270. (Visited on 02/23/2021) (cit. on pp. 23, 100, 105, 131).

Bojanowski, P., E. Grave, A. Joulin, T. Mikolov (2017). 'Enriching Word Vectors with Subword Information'. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. (Visited on 10/23/2019) (cit. on pp. 30, 67, 81, 104, 105).

Bollegala, D., K. Hayashi, K.-i. Kawarabayashi (July 2018). 'Think Globally, Embed Locally — Locally Linear Meta-embedding of Words'. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, pp. 3970–3976. (Visited on 02/20/2019) (cit. on pp. 69, 143).

Borchert, F., C. Lohr, L. Modersohn, T. Langer, M. Follmann, J. P. Sachs, U. Hahn, M.-P. Schapranow (Nov. 2020). 'GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines'. In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Online: Association for Computational Linguistics, pp. 38–48. arXiv: 2007.06400. (Visited on 12/09/2022) (cit. on pp. 5, 129).

Borgeaud, S., A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. V. D. Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. D. L. Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, L. Sifre (June 2022). 'Improving Language Models by Retrieving from Trillions of Tokens'. In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, pp. 2206–2240. (Visited on 11/05/2023) (cit. on p. 145).

Born, J., D. Beymer, D. Rajan, A. Coy, V. V. Mukherjee, M. Manica, P. Prasanna, D. Ballah, M. Guindy, D. Shaham (2021). 'On the Role of Artificial Intelligence in Medical Imaging of Covid-19'. In: *Patterns* (cit. on p. 20).

Bouma, G. (2009). 'Normalized (Pointwise) Mutual Information in Collocation Extraction'. In: *Proceedings of GSCL*, pp. 31–40 (cit. on pp. 95, 113, 114).

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei (July 2020). 'Language Models Are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems*. Vol. 33. Online: Curran Associates, Inc., pp. 1877–1901. arXiv: 2005.14165. (Visited on 02/23/2021) (cit. on pp. 26, 30, 31, 34, 103, 128).

Bunescu, R., M. Paşca (Apr. 2006). 'Using Encyclopedic Knowledge for Named Entity Disambiguation'. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics. (Visited on 02/24/2021) (cit. on p. 104).

Catling, F., G. P. Spithourakis, S. Riedel (Dec. 2018). 'Towards Automated Clinical Coding'. In: *International Journal of Medical Informatics* 120, pp. 50–61. (Visited on 05/31/2021) (cit. on pp. 21, 130).

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil (Mar. 2018). 'Universal Sentence Encoder'. In: *arXiv:1803.11175 [cs]*. arXiv: 1803.11175 [cs]. (Visited on 09/17/2018) (cit. on p. 78).

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan (Oct. 2001). 'A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries'. In: *Journal of Biomedical Informatics* 34.5, pp. 301–310. (Visited on 02/23/2021) (cit. on p. 131).

Chapman, W. W., P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, O. Uzuner (2011 Sep-Oct). 'Overcoming Barriers to NLP for Clinical Text: The Role of Shared Tasks and the Need for Additional Creative Solutions'. In: *Journal of the American Medical Informatics Association: JAMIA* 18.5, pp. 540–543 (cit. on p. 104).

Chelba, C. (June 2010). 'Statistical Language Modeling'. In: *The Handbook of Computational Linguistics and Natural Language Processing*. Ed. by A. Clark, C. Fox, S. Lappin. Oxford, UK: Wiley-Blackwell, pp. 74–104. (Visited on 11/04/2019) (cit. on p. 67).

Chen, M., Z. Chu, Y. Chen, K. Stratos, K. Gimpel (2019). 'EntEval: A Holistic Evaluation Benchmark for Entity Representations'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by K. Inui, J. Jiang, V. Ng, X. Wan. Association for Computational Linguistics, pp. 421–433. arXiv: 1909.00137 (cit. on pp. 33, 143).

Chiticariu, L., R. Krishnamurthy, Y. Li, S. Raghavan, F. R. Reiss, S. Vaithyanathan (2010). 'SystemT: An Algebraic Approach to Declarative Information Extraction'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 128–137. (Visited on 04/03/2012) (cit. on p. 7).

Chiticariu, L., Y. Li, F. R. Reiss (2013). 'Rule-Based Information Extraction Is Dead! Long Live Rule-Based Information Extraction Systems!' In: *EMNLP*, pp. 827–832. (Visited on 05/03/2016) (cit. on p. 24).

Choi, Y., C. Y.-I. Chiu, D. Sontag (July 2016). 'Learning Low-Dimensional Representations of Medical Concepts'. In: *AMIA Summits on Translational Science Proceedings* 2016, pp. 41–50. (Visited on 02/03/2021) (cit. on pp. 33, 131).

Christensen, J., S. Soderland, O. Etzioni, et al. (2011). 'An Analysis of Open Information Extraction Based on Semantic Role Labeling'. In: *Proceedings of the Sixth International Conference on Knowledge Capture*. ACM, pp. 113–120. (Visited on 08/28/2017) (cit. on p. 25).

Coates, J., D. Bollegala (2018). 'Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Asso-

ciation for Computational Linguistics, pp. 194–198. (Visited on 02/20/2019) (cit. on pp. 34, 69, 73, 143).

Codd, E. F. (June 1970). 'A Relational Model of Data for Large Shared Data Banks'. In: *Communications of the ACM* 13.6, pp. 377–387. (Visited on 05/11/2021) (cit. on pp. 60, 112).

Conneau, A., D. Kiela (2018). 'SentEval: An Evaluation Toolkit for Universal Sentence Representations'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Language Resource Association. (Visited on 02/13/2019) (cit. on pp. 12, 61, 63, 64, 75, 76).

Conneau, A., G. Kruszewski, G. Lample, L. Barrault, M. Baroni (2018). 'What You Can Cram into a Single \$&!#* Vector: Probing Sentence Embeddings for Linguistic Properties'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. (Visited on 09/03/2018) (cit. on p. 33).

Connor, J. A., R. Thiagarajan (May 2007). 'Hypoplastic Left Heart Syndrome'. In: *Orphanet Journal of Rare Diseases* 2, p. 23. (Visited on 10/15/2023) (cit. on p. 72).

Cotik, V., R. Roller, F. Xu, H. Uszkoreit, K. Budde, D. Schmidt (Dec. 2016). 'Negation Detection in Clinical Reports Written in German'. In: *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 115–124. (Visited on 02/04/2021) (cit. on p. 131).

Croskerry, P. (2009). 'A Universal Model of Diagnostic Reasoning'. In: *Academic medicine* 84.8, pp. 1022–1028 (cit. on pp. 21, 100, 101, 130).

Cui, L., F. Wei, M. Zhou (July 2018). 'Neural Open Information Extraction'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 407–413. (Visited on 11/29/2021) (cit. on p. 26).

Cui, L., A. Bozorgi, S. D. Lhatoo, G.-Q. Zhang, S. S. Sahoo (Nov. 2012). 'EpiDEA: Extracting Structured Epilepsy and Seizure Information from Patient Discharge Summaries for Cohort Identification'. In: *AMIA Annual Symposium Proceedings* 2012, pp. 1191–1200. (Visited on 05/19/2020) (cit. on pp. 1, 6, 7, 19, 20, 119, 130, 132).

D'Souza, J., V. Ng (2015). 'Sieve-Based Entity Linking for the Biomedical Domain'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 297–302. (Visited on 07/04/2019) (cit. on p. 131).

De Fauw, J., J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin (2018). 'Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease'. In: *Nature medicine* 24.9, pp. 1342–1350 (cit. on pp. 1, 19, 21, 130).

De Marneffe, M.-C., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, C. D. Manning (2014). 'Universal Stanford Dependencies: A Cross-Linguistic Typology.' In: *LREC*. Vol. 14, pp. 4585–92. (Visited on 08/10/2016) (cit. on p. 23).

Del Corro, L. (2016). 'Methods for Open Information Extraction and Sense Disambiguation on Natural Language Text'. PhD thesis. Saarbrücken: Universität des Saarlandes. (Visited on 06/16/2016) (cit. on p. 25).

Del Corro, L., R. Gemulla (2013). 'Clausie: Clause-Based Open Information Extraction'. In: *Proceedings of the 22nd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 355–366. (Visited on 06/02/2016) (cit. on pp. 24, 25, 41, 47, 50, 58).

Demner-Fushman, D., W. W. Chapman, C. J. McDonald (Oct. 2009). 'What Can Natural Language Processing Do for Clinical Decision Support?' In: *Journal of Biomedical Informatics*. Biomedical Natural Language Processing 42.5, pp. 760–772. (Visited on 10/31/2016) (cit. on pp. 1, 21, 64).

Devlin, J., M.-W. Chang, K. Lee, K. Toutanova (2019). '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, Volume 1 (Long and Short Papers)*, pp. 4171–4186. arXiv: 1810.04805 (cit. on pp. 8, 15, 22, 25, 26, 30–32, 34, 63, 65–67, 74, 128, 143).

Dieng, A. B., C. Wang, J. Gao, J. W. Paisley (Apr. 2017). 'TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency'. In: *Conference Track Proceedings of the 5th International Conference on Learning Representations {ICLR}*. Toulon, France: OpenReview.net (cit. on p. 33).

Dolan, B., C. Quirk, C. Brockett (2004). 'Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources'. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, p. 350 (cit. on p. 75).

Doshi-Velez, F., Y. Ge, I. Kohane (Jan. 2014). 'Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis'. In: *Pediatrics* 133.1, e54–63 (cit. on p. 19).

Dudley, J. J., P. O. Kristensson (2018). 'A Review of User Interface Design for Interactive Machine Learning'. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.2, pp. 1–37 (cit. on p. 114).

Eiband, M., H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, H. Hussmann (2018). 'Bringing Transparency Design into Practice'. In: *23rd International Conference on Intelligent User Interfaces*. IUI '18. New York, NY, USA: ACM, pp. 211–223. (Visited on 08/30/2018) (cit. on p. 114).

Elasticsearch (Feb. 2021). elasticsearch B.V. (Visited on 02/26/2021) (cit. on p. 105).

Ely, J. W., J. A. Osheroff, P. N. Gorman, M. H. Ebell, M. L. Chambliss, E. A. Pifer, P. Z. Stavri (Aug. 2000). 'A Taxonomy of Generic Clinical Questions: Classification Study'. In: *BMJ* 321.7258, pp. 429–432. (Visited on 05/17/2018) (cit. on p. 3).

Eslami, S., N. F. de Keizer, D. A. Dongelmans, E. de Jonge, M. J. Schultz, A. Abu-Hanna (Jan. 2012). 'Effects of Two Different Levels of Computerized Decision Support on Blood Glucose Regulation in Critically Ill Patients'. In: *International Journal of Medical Informatics* 81.1, pp. 53–60 (cit. on pp. 1, 19, 130).

Esteva, A., K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher (Jan. 2021). 'Deep Learning-Enabled Medical Computer Vision'. In: *npj Digital Medicine* 4.1, pp. 1–9. (Visited on 08/07/2022) (cit. on pp. 4, 7, 20).

Ethayarajh, K. (2019). 'How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 55–65. (Visited on 11/21/2019) (cit. on p. 33).

Etzioni, O., A. Fader, J. Christensen, S. Soderland, M. Mausam (2011). 'Open Information Extraction: The Second Generation.' In: *IJCAI*. Vol. 11, pp. 3–10. (Visited on 06/02/2016) (cit. on pp. 24, 25).

Fader, A., S. Soderland, O. Etzioni (2011). 'Identifying Relations for Open Information Extraction'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1535–1545. (Visited on 06/02/2016) (cit. on pp. 25, 51).

Fader, A., L. Zettlemoyer, O. Etzioni (2013). 'Paraphrase-Driven Learning for Open Question Answering'. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1608–1618 (cit. on p. 24).

Faviez, C., X. Chen, N. Garcelon, A. Neuraz, B. Knebelmann, R. Salomon, S. Lyonnet, S. Saunier, A. Burgun (Dec. 2020). 'Diagnosis Support Systems for Rare Diseases:

A Scoping Review'. In: *Orphanet Journal of Rare Diseases* 15.1, p. 94. (Visited on 07/15/2020) (cit. on p. 1).

Fellbaum, C., ed. (May 1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA, USA: A Bradford Book (cit. on p. 43).

Finlayson, S. G., P. LePendu, N. H. Shah (Dec. 2014). 'Building the Graph of Medicine from Millions of Clinical Narratives'. In: *Scientific Data* 1.1, p. 140032. (Visited on 02/22/2021) (cit. on p. 25).

Firth, J. R. (1957). 'A Synopsis of Linguistic Theory 1930-55. (Special Volume of the Philological Society)'. In: *Studies in Linguistic Analysis* 1952–59, pp. 1–32 (cit. on p. 27).

Firth, J. R. (1961). *Papers in Linguistics 1934-1951*. Oxford University Press (cit. on p. 27).

Fu, S., D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen (2020). 'Clinical Concept Extraction: A Methodology Review'. In: *Journal of Biomedical Informatics* 109, p. 103526 (cit. on pp. 3, 6, 23, 105, 131).

Gamble, P., T. Jaroensri, H. Wang, F. Tan, M. Moran, T. Brown, I. Flament, E. A. Rakha, M. Toss, D. J. Dabbs, P. Regitnig, N. Olson, J. H. Wren, C. Robinson, G. Corrado, L. Peng, Y. Liu, C. Mermel, D. Steiner, C. Chen (2021). 'Determining Breast Cancer Biomarker Status and Associated Morphological Features Using Deep Learning'. In: *Nature Communications Medicine*. (Visited on 08/25/2021) (cit. on p. 20).

Gashteovski, K. (2020). 'Compact Open Information Extraction: Methods, Corpora, Analysis'. PhD thesis. Mannheim: University of Mannheim (cit. on pp. 24, 25).

Gashteovski, K., R. Gemulla, L. del Corro (2017). 'Minie: Minimizing Facts in Open Information Extraction'. In: Association for Computational Linguistics (cit. on p. 25).

Georgiou, A., M. Prgomet, A. Markewycz, E. Adams, J. I. Westbrook (2011). 'The Impact of Computerized Provider Order Entry Systems on Medical-Imaging Services: A Systematic Review'. In: *Journal of the American Medical Informatics Association : JAMIA* 18.3, pp. 335–340. (Visited on 08/25/2021) (cit. on p. 20).

Gers, F. A., J. A. Pérez-Ortiz, D. Eck, J. Schmidhuber (2002). 'Learning Context Sensitive Languages with LSTM Trained with Kalman Filters'. In: *Artificial Neural Networks — ICANN 2002*. Ed. by J. R. Dorronsoro. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 655–660 (cit. on p. 88).

Gers, F. A., J. A. Schmidhuber, F. A. Cummins (Oct. 2000). 'Learning to Forget: Continual Prediction with LSTM'. In: *Neural computation* 12.10, pp. 2451–2471. (Visited on 06/15/2018) (cit. on pp. 32, 68, 93, 131).

Giardino, A., S. Gupta, E. Olson, K. Sepulveda, L. Lenchik, J. Ivanidze, R. Rakow-Penner, M. J. Patel, R. M. Subramaniam, D. Ganeshan (May 2017). 'Role of Imaging in the Era of Precision Medicine'. In: *Academic Radiology* 24.5, pp. 639–649 (cit. on p. 20).

Gillick, D., S. Kulkarni, L. Lansing, A. Presta, J. Baldridge, E. Ie, D. Garcia-Olano (2019). 'Learning Dense Representations for Entity Retrieval'. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 528–537. arXiv: 1909.10506. (Visited on 12/09/2022) (cit. on pp. 63, 68).

Glicksberg, B. S., R. Miotto, K. W. Johnson, K. Shameer, L. Li, R. Chen, J. T. Dudley (Jan. 2018). 'Automated Disease Cohort Selection Using Word Embeddings from Electronic Health Records'. In: *Biocomputing 2018*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, pp. 145–156. (Visited on 03/22/2021) (cit. on pp. 6, 8, 20, 98, 108, 128, 132).

Glicksberg, B. S., B. Oskotsky, P. M. Thangaraj, N. Giangreco, M. A. Badgeley, K. W. Johnson, D. Datta, V. A. Rudrapatna, N. Rappoport, M. M. Shervey, R. Miotto, T. C. Goldstein, E. Rutenberg, R. Frazier, N. Lee, S. Israni, R. Larsen, B. Percha, L. Li, J. T. Dudley, N. P. Tatonetti, A. J. Butte (Nov. 2019). 'PatientExploreR: An Extensible Application for Dynamic Visualization of Patient Clinical History from Electronic Health Records in the OMOP Common Data Model'. In: *Bioinformatics* 35.21, pp. 4515–4518. (Visited on 02/15/2021) (cit. on p. 108).

Glorot, X., Y. Bengio (May 2010). 'Understanding the Difficulty of Training Deep Feedforward Neural Networks'. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh, M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256 (cit. on p. 71).

Glorot, X., A. Bordes, Y. Bengio (Apr. 2011). 'Deep Sparse Rectifier Neural Networks'. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 315–323 (cit. on p. 71).

Goldenberg, S. L., G. Nir, S. E. Salcudean (July 2019). 'A New Era: Artificial Intelligence and Machine Learning in Prostate Cancer'. In: *Nature Reviews Urology* 16.7, pp. 391–403. (Visited on 05/31/2021) (cit. on pp. 1, 21, 130).

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio (2014). 'Generative Adversarial Nets'. In: *Advances in Neural Information Processing Systems*, pp. 2672–2680 (cit. on p. 20).

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Vol. 385. Berlin Heidelberg: Springer. (Visited on 09/08/2015) (cit. on pp. 68, 93).

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon (2021). 'Domain-Specific Language Model Pretraining for Biomedical Natural Language

Processing'. In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1, pp. 1–23. arXiv: 2007.15779 (cit. on pp. 6, 8, 31–33, 103, 128, 129).

Harris, Z. S. (Aug. 1954). 'Distributional Structure'. In: *WORD* 10.2-3, pp. 146–162. (Visited on 02/23/2017) (cit. on pp. 27, 28).

Hearst, M. A. (1992). 'Automatic Acquisition of Hyponyms from Large Text Corpora'. In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. (Visited on 11/07/2021) (cit. on p. 43).

Heilman, J. M., A. G. West (Mar. 2015). 'Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language'. In: *Journal of Medical Internet Research* 17.3. (Visited on 02/15/2019) (cit. on p. 104).

Heinz, S., C. Bracher, R. Vollgraf (2017). 'An LSTM-Based Dynamic Customer Model for Fashion Recommendation'. In: *Proceedings of the 1st Workshop on Temporal Reasoning in Recommender Systems Co-Located with 11th International Conference on Recommender Systems (RecSys 2017)*. Vol. 1922. Como, Italy: CEUR-WS, p. 5 (cit. on p. 69).

Hochreiter, S., J. Schmidhuber (1997). 'Long Short-Term Memory'. In: *Neural Comput.* 9.8, pp. 1735–1780 (cit. on pp. 88, 92, 131).

Hong, N., A. Wen, F. Shen, S. Sohn, S. Liu, H. Liu, G. Jiang (May 2018). 'Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data'. In: *AMIA Summits on Translational Science Proceedings* 2018, pp. 74–83. (Visited on 04/09/2019) (cit. on p. 2).

Howard, J., S. Ruder (2018). 'Universal Language Model Fine-tuning for Text Classification'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 328–339. (Visited on 02/25/2019) (cit. on p. 107).

Hu, M., B. Liu (2004). 'Mining and Summarizing Customer Reviews'. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 168–177 (cit. on p. 75).

Huang, Z., W. Dong, L. Ji, C. Gan, X. Lu, H. Duan (Feb. 2014). 'Discovery of Clinical Pathway Patterns from Event Logs Using Probabilistic Topic Models'. In: *Journal of Biomedical Informatics* 47, pp. 39–57. (Visited on 05/25/2020) (cit. on pp. 2, 18, 19).

Huang, Z., Z. Ge, W. Dong, K. He, H. Duan (2018). 'Probabilistic Modeling Personalized Treatment Pathways Using Electronic Health Records'. In: *Journal of biomedical informatics* 86, pp. 33–48 (cit. on p. 19).

Huang, Z., W. Xu, K. Yu (2015). 'Bidirectional LSTM-CRF Models for Sequence Tagging'. In: *arXiv preprint arXiv:1508.01991*. arXiv: 1508.01991. (Visited on 09/28/2015) (cit. on p. 32).

Huhtala, M., S. Geurts, S. Mauno, T. Feldt (2021). 'Intensified Job Demands in Healthcare and Their Consequences for Employee Well-Being and Patient Satisfaction: A Multilevel Approach'. In: *Journal of Advanced Nursing* 77.9, pp. 3718–3732. (Visited on 08/07/2022) (cit. on p. 3).

Jadhav, A., K. C. Wong, J. T. Wu, M. Moradi, T. Syeda-Mahmood (2020). 'Combining Deep Learning and Knowledge-driven Reasoning for Chest X-Ray Findings Detection'. In: *AMIA Annual Symposium Proceedings*. Vol. 2020. American Medical Informatics Association, p. 593 (cit. on p. 20).

Jain, A., M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, J. Baldridge (Nov. 2021). 'MURAL: Multimodal, Multitask Representations across Languages'. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3449–3463. arXiv: 2109.05125 (cit. on p. 34).

Jauregi Unanue, I., E. Zare Borzeshi, M. Piccardi (Dec. 2017). 'Recurrent Neural Networks with Specialized Word Embeddings for Health-Domain Named-Entity Recognition'. In: *Journal of Biomedical Informatics* 76, pp. 102–109. (Visited on 02/16/2021) (cit. on pp. 4, 23, 104, 131).

Ji, Y., C. Tan, S. Martschat, Y. Choi, N. A. Smith (Aug. 2017). 'Dynamic Entity Representations in Neural Language Models'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1830–1839. arXiv: 1708.00781. (Visited on 11/21/2019) (cit. on p. 33).

Jia, C., Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig (Feb. 2021). 'Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision'. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 38, pp. 4904–4916. arXiv: 2102.05918. (Visited on 04/22/2021) (cit. on p. 98).

Jiang, D., J. Liao, H. Duan, Q. Wu, G. Owen, C. Shu, L. Chen, Y. He, Z. Wu, D. He, W. Zhang, Z. Wang (June 2020). 'A Machine Learning-Based Prognostic Predictor for Stage III Colon Cancer'. In: *Scientific Reports* 10.1, p. 10333. (Visited on 05/31/2021) (cit. on pp. 1, 21, 130).

Jiang, T., Q. Zeng, T. Zhao, B. Qin, T. Liu, N. V. Chawla, M. Jiang (May 2021). 'Biomedical Knowledge Graphs Construction From Conditional Statements'. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.3, pp. 823–835 (cit. on p. 38).

Jiang, T., T. Zhao, B. Qin, T. Liu, N. V. Chawla, M. Jiang (June 2020). 'Canonicalizing Open Knowledge Bases with Multi-Layered Meta-Graph Neural Network'. In: *arXiv:2006.09610 [cs]*. arXiv: 2006.09610 [cs]. (Visited on 11/29/2021) (cit. on p. 26).

Johnson, A. E. W., T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark (May 2016). 'MIMIC-III, a Freely Accessible Critical Care Database'. In: *Scientific Data* 3.1, pp. 1–9. (Visited on 05/18/2020) (cit. on pp. 5, 100, 118, 129).

Joulin, A., E. Grave, P. Bojanowski, T. Mikolov (2017). 'Bag of Tricks for Efficient Text Classification'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Vol. 2. Valencia, Spain, pp. 427–431. (Visited on 12/08/2016) (cit. on pp. 63, 67).

Jungmann, S. M., T. Klan, S. Kuhn, F. Jungmann (Oct. 2019). 'Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users'. In: *JMIR Formative Research* 3.4. (Visited on 05/30/2021) (cit. on pp. 19, 130).

Jurafsky, D., H. M. James (Sept. 2021). *Speech and Language Processing*. Third Edition draft. (Visited on 11/29/2021) (cit. on pp. 7, 22, 23).

Kadry, A., L. Dietz (2017). 'Open Relation Extraction for Support Passage Retrieval: Merit and Open Issues'. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1149–1152 (cit. on p. 25).

Khot, T., A. Sabharwal, P. Clark (July 2017). 'Answering Complex Questions Using Open Information Extraction'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 311–316. arXiv: 1704.05572 (cit. on p. 24).

Kiela, D., C. Wang, K. Cho (2018). 'Dynamic Meta-Embeddings for Improved Sentence Representations'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1466–1477 (cit. on pp. 33, 34, 69, 73, 143).

Kilias, T., A. Löser, P. Andritsos (2015). 'INDREX: In-Database Relation Extraction'. In: *Information Systems* 53, pp. 124–144 (cit. on pp. 7, 23, 40, 41).

Kingma, D., J. Ba (2015). 'ADAM: A Method for Stochastic Optimization'. In: *ICLR'15*. (Visited on 06/30/2017) (cit. on pp. 71, 94).

Kinsman, L., T. Rotter, E. James, P. Snow, J. Willis (May 2010). 'What Is a Clinical Pathway? Development of a Definition to Inform the Debate'. In: *BMC Medicine* 8.1, p. 31. (Visited on 05/22/2020) (cit. on p. 18).

Kiros, R., Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler (2015). 'Skip-Thought Vectors'. In: *Proceedings of the 28th International Conference on Neural*

Information Processing Systems - Volume 2. NIPS'15. Cambridge, MA, USA: MIT Press, pp. 3276–3284. (Visited on 04/05/2016) (cit. on pp. 31, 32, 67).

Knake, L. A., M. Ahuja, E. L. McDonald, K. K. Ryckman, N. Weathers, T. Burstain, J. M. Dagle, J. C. Murray, P. Nadkarni (2016). 'Quality of EHR Data Extractions for Studies of Preterm Birth in a Tertiary Care Center: Guidelines for Obtaining Reliable Data'. In: *BMC pediatrics* 16.1, pp. 1–8 (cit. on p. 18).

Köhn, A. (2015). 'What's in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation'. In: (visited on 09/25/2015) (cit. on p. 33).

Kolluru, K., V. Adlakha, S. Aggarwal, Mausam, S. Chakrabarti (Nov. 2020). 'OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3748–3761. arXiv: 2010.03147 (cit. on p. 26).

Kolluru, K., S. Aggarwal, V. Rathore, Mausam, S. Chakrabarti (July 2020). 'IMoJIE: Iterative Memory-Based Joint Open Information Extraction'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5871–5886. arXiv: 2005.08178 (cit. on p. 26).

Koopman, B., L. Cripwell, G. Zuccon (Aug. 2017). 'Generating Clinical Queries from Patient Narratives: A Comparison between Machines and Humans'. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. New York, NY, USA: Association for Computing Machinery, pp. 853–856. (Visited on 03/01/2021) (cit. on p. 21).

Kramer, M. A. (1991). 'Nonlinear Principal Component Analysis Using Autoassociative Neural Networks'. In: *AIChE Journal* 37.2, pp. 233–243. (Visited on 05/31/2021) (cit. on p. 132).

Krishnamurthy, R., Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, H. Zhu (Mar. 2009). 'SystemT: A System for Declarative Information Extraction'. In: *SIGMOD Rec.* 37.4, pp. 7–13. (Visited on 04/26/2012) (cit. on pp. 23, 38, 39, 45).

Kudo, T., J. Richardson (Nov. 2018). 'SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. arXiv: 1808.06226 (cit. on p. 22).

Kuebler, J., L. Tong, M. Jiang (Aug. 2021). 'Multi-Round Parsing-based Multiword Rules for Scientific OpenIE'. In: *arXiv:2108.02074 [cs]*. arXiv: 2108.02074 [cs]. (Visited on 08/23/2021) (cit. on p. 38).

Landi, I., B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, R. Miotto (July 2020). 'Deep Representation Learning of Electronic Health Records to Unlock Patient Stratification at Scale'. In: *npj Digital Medicine* 3.1, pp. 1–11. (Visited on 03/25/2021) (cit. on pp. 2, 8, 19, 20, 111, 128, 132, 143, 144).

Le, Q. V., T. Mikolov (2014). 'Distributed Representations of Sentences and Documents.' In: *ICML'14*. Vol. 32, pp. 1188–1196. (Visited on 10/12/2016) (cit. on pp. 30, 68, 91).

Leaman, R., R. Khare, Z. Lu (Oct. 2015). 'Challenges in Clinical Natural Language Processing for Automated Disorder Normalization'. In: *Journal of Biomedical Informatics* 57, pp. 28–37. (Visited on 10/31/2019) (cit. on pp. 2, 60, 64, 104, 131).

Leaman, R., Z. Lu (Sept. 2016). 'TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models'. In: *Bioinformatics* 32.18, pp. 2839–2846. (Visited on 02/23/2021) (cit. on p. 131).

Lecun, Y., L. Bottou, Y. Bengio, P. Haffner (Nov. 1998). 'Gradient-Based Learning Applied to Document Recognition'. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 (cit. on pp. 131, 132).

LeCun, Y., Y. Bengio (1995). 'Convolutional Networks for Images, Speech, and Time Series'. In: *The handbook of brain theory and neural networks* 3361.10, p. 1995 (cit. on p. 20).

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang (Sept. 2019). 'BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining'. In: *Bioinformatics* (cit. on pp. 5, 7, 21, 31, 32, 67, 68, 83, 103, 128).

Lee, J., D. Maslove, J. A. Dubin (2015). 'Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric'. In: *PloS one* (cit. on pp. 108, 130, 132).

Levandowsky, M., D. Winter (Nov. 1971). 'Distance between Sets'. In: *Nature* 234.5323, pp. 34–35. (Visited on 04/14/2021) (cit. on p. 110).

Li, L., W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, J. T. Dudley (Oct. 2015). 'Identification of Type 2 Diabetes Subgroups through Topological Analysis of Patient Similarity'. In: *Science Translational Medicine* 7.311, 311ra174–311ra174. (Visited on 05/22/2020) (cit. on pp. 6, 18–20, 130, 132).

Li, Q., X. Wang, Y. Zhang, F. Ling, C. H. Wu, J. Han (Dec. 2018). 'Pattern Discovery for Wide-Window Open Information Extraction in Biomedical Literature'. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 420–427 (cit. on p. 25).

Li, W. T., J. Ma, N. Shende, G. Castaneda, J. Chakladar, J. C. Tsai, L. Apostol, C. O. Honda, J. Xu, L. M. Wong, T. Zhang, A. Lee, A. Gnanasekar, T. K. Honda, S. Z. Kuo, M. A. Yu, E. Y. Chang, M. ' R. Rajasekaran, W. M. Ongkeko (Sept. 2020). 'Using Machine Learning

of Clinical Data to Diagnose COVID-19: A Systematic Review and Meta-Analysis'. In: *BMC Medical Informatics and Decision Making* 20.1, p. 247. (Visited on 01/08/2021) (cit. on p. 20).

Likert, R. (1932). 'A Technique for the Measurement of Attitudes'. In: *Archives of Psychology* 22 140, pp. 55–55 (cit. on p. 121).

Lin, X., H. Li, H. Xin, Z. Li, L. Chen (Mar. 2020). 'KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking'. In: *Proceedings of the VLDB Endowment* 13.7, pp. 1035–1049. (Visited on 11/29/2021) (cit. on p. 25).

Ling, J., N. FitzGerald, Z. Shan, L. B. Soares, T. Févry, D. Weiss, T. Kwiatkowski (Jan. 2020). 'Learning Cross-Context Entity Representations from Text'. In: *arXiv:2001.03765 [Cs]*. arXiv: 2001.03765 [cs]. (Visited on 01/22/2020) (cit. on p. 33).

Liu, H., S. J. Bielinski, S. Sohn, S. Murphy, K. B. Wagholikar, S. R. Jonnalagadda, K. Ravikumar, S. T. Wu, I. J. Kullo, C. G. Chute (2013 -3- 18). 'An Information Extraction Framework for Cohort Identification Using Electronic Health Records'. In: *AMIA Summits on Translational Science Proceedings* 2013, pp. 149–153. (Visited on 05/19/2020) (cit. on p. 4).

Liu, J., X. Ren, J. Shang, T. Cassidy, C. R. Voss, J. Han (2016). 'Representing Documents via Latent Keyphrase Inference'. In: *WWW'16*, pp. 1057–1067 (cit. on p. 33).

Liu, R., R. V. Srinivasan, K. Zolfaghar, S.-C. Chin, S. B. Roy, A. Hasan, D. Hazel (2014). 'Pathway-Finder: An Interactive Recommender System for Supporting Personalized Care Pathways'. In: *2014 IEEE International Conference on Data Mining Workshop*. IEEE, pp. 1219–1222 (cit. on p. 19).

Liu, Y., A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. d. O. Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. Huang, Y. Liu, R. C. Dunn, D. Coz (Sept. 2019). 'A Deep Learning System for Differential Diagnosis of Skin Diseases'. In: *Nature Medicine* 26.6, pp. 900–908. (Visited on 04/14/2020) (cit. on pp. 1, 130).

MacAvaney, S., A. Yates, A. Cohan, L. Soldaini, K. Hui, N. Goharian, O. Frieder (June 2018). 'Characterizing Question Facets for Complex Answer Retrieval'. In: *SIGIR '18 The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Ann Arbor, MI, USA: ACM, pp. 1205–1208. arXiv: 1805.00791. (Visited on 02/21/2019) (cit. on p. 68).

Mahoney, C. D., C. M. Berard-Collins, R. Coleman, J. F. Amaral, C. M. Cotter (Sept. 2007). 'Effects of an Integrated Clinical Information System on Medication Safety in a Multi-Hospital Setting'. In: *American journal of health-system pharmacy: AJHP: official journal*

of the American Society of Health-System Pharmacists 64.18, pp. 1969–1977 (cit. on pp. 1, 19, 130).

Mandel, S. (2019). *Almondtools/Stringsearchalgorithms*. (Visited on 02/24/2021) (cit. on p. 105).

Manning, C. D., P. Raghavan, H. Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press (cit. on p. 22).

Manning, C., H. Schutze (1999). *Foundations of Statistical Natural Language Processing*. MIT press (cit. on pp. 7, 24).

Marcus, M. P., B. Santorini, M. A. Marcinkiewicz (1993). 'Building a Large Annotated Corpus of English: The Penn Treebank'. In: *Computational Linguistics* 19.2, pp. 313–330. (Visited on 11/29/2021) (cit. on pp. 7, 22).

Marelli, M., S. Menini, M. Baroni, L. Bentivogli, R. bernardi, R. Zamparelli (2014). 'A SICK Cure for the Evaluation of Compositional Distributional Semantic Models'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA). (Visited on 02/28/2019) (cit. on p. 75).

Mausam, M. (July 2016). 'Open Information Extraction Systems and Downstream Applications'. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. New York. (Visited on 09/06/2016) (cit. on pp. 24, 25).

Mausam, M. Schmitz, S. Soderland, R. Bart, O. Etzioni (2012). 'Open Language Learning for Information Extraction'. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534 (cit. on p. 25).

McEvoy, D. S., D. F. Sittig, T.-T. Hickman, S. Aaron, A. Ai, M. Amato, D. W. Bauer, G. M. Fraser, J. Harper, A. Kennemer, M. A. Krall, C. U. Lehmann, S. Malhotra, D. R. Murphy, B. O'Kelley, L. Samal, R. Schreiber, H. Singh, E. J. Thomas, C. V. Vartian, J. Westmorland, A. B. McCoy, A. Wright (Mar. 2017). 'Variation in High-Priority Drug-Drug Interaction Alerts across Institutions and Electronic Health Records'. In: *Journal of the American Medical Informatics Association : JAMIA* 24.2, pp. 331–338. (Visited on 05/31/2021) (cit. on pp. 1, 19, 130).

McMullin, S. T., T. P. Lonergan, C. S. Rynearson, T. D. Doerr, P. A. Veregge, E. S. Scanlan (2004 Sep-Oct). 'Impact of an Evidence-Based Computerized Decision Support System on Primary Care Prescription Costs'. In: *Annals of Family Medicine* 2.5, pp. 494–498 (cit. on pp. 1, 19, 130).

Mehrabi, S., A. Krishnan, S. Sohn, A. M. Roch, H. Schmidt, J. Kesterson, C. Beesley, P. Dexter, C. Max Schmidt, H. Liu, M. Palakal (Apr. 2015). 'DEEPEN: A Negation

Detection System for Clinical Text Incorporating Dependency Relation into NegEx'. In: *Journal of Biomedical Informatics* 54, pp. 213–219. (Visited on 02/03/2021) (cit. on p. 131).

Mesquita, F., J. Schmidek, D. Barbosa (2013). 'Effectiveness and Efficiency of Open Relation Extraction'. In: *EMNLP'13*. Association for Computational Linguistics, pp. 447–457. (Visited on 08/22/2016) (cit. on pp. xv, xvi, 48, 49, 52).

Michael, T., A. Akbik (2015). 'SCHNAPPER: A Web Toolkit for Exploratory Relation Extraction'. In: *ACL System Demonstrations* (cit. on p. 40).

Mikolov, T., K. Chen, G. Corrado, J. Dean (2013). 'Efficient Estimation of Word Representations in Vector Space'. In: *1st International Conference on Learning Representations, Workshop Track Proceedings*. Scottsdale, Arizona, USA (cit. on pp. 25, 30, 32, 63, 132).

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin (May 2018). 'Advances in Pre-Training Distributed Word Representations'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). (Visited on 10/23/2019) (cit. on pp. 65–67, 71, 73).

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, J. Dean (2013). 'Distributed Representations of Words and Phrases and Their Compositionality'. In: *NIPS'13*, pp. 3111–3119. (Visited on 07/04/2016) (cit. on pp. 67, 132).

Mikolov, T., W.-t. Yih, G. Zweig (2013). 'Linguistic Regularities in Continuous Space Word Representations.' In: *HLT-NAACL*, pp. 746–751. (Visited on 08/06/2015) (cit. on p. 30).

Miniukovich, A., S. Sulpizio, A. De Angeli (2018). 'Visual Complexity of Graphical User Interfaces'. In: *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, pp. 1–9 (cit. on p. 114).

Mintz, M., S. Bills, R. Snow, D. Jurafsky (2009). 'Distant Supervision for Relation Extraction Without Labeled Data'. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. ACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1003–1011. (Visited on 06/12/2019) (cit. on p. 71).

Miotto, R., L. Li, B. A. Kidd, J. T. Dudley (May 2016). 'Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records'. In: *Scientific Reports* 6.1, pp. 1–10. (Visited on 05/18/2020) (cit. on pp. 8, 19, 20, 128, 132, 144).

Miotto, R., F. Wang, S. Wang, X. Jiang, J. T. Dudley (Nov. 2018). 'Deep Learning for Healthcare: Review, Opportunities and Challenges'. In: *Briefings in Bioinformatics* 19.6, pp. 1236–1246. (Visited on 03/25/2021) (cit. on pp. 19, 98, 111, 128, 130, 144).

Mohan, S., D. Li (2019). 'MedMentions: A Large Biomedical Corpus Annotated with {UMLS} Concepts'. In: *Automated Knowledge Base Construction (AKBC)* (cit. on p. 104).

Morgan, A. A., L. Hirschman, M. Colosimo, A. S. Yeh, J. B. Colombe (Dec. 2004). 'Gene Name Identification and Normalization Using a Model Organism Database'. In: *Journal of Biomedical Informatics*. Named Entity Recognition in Biomedicine 37.6, pp. 396–410. (Visited on 09/25/2019) (cit. on p. 71).

Mroueh, Y., E. Marcheret, V. Goel (Nov. 2015). 'Asymmetrically Weighted CCA And Hierarchical Kernel Sentence Embedding For Image & Text Retrieval'. In: *arXiv:1511.06267 [cs]*. arXiv: 1511.06267 [cs]. (Visited on 06/12/2018) (cit. on p. 34).

Mueller, D., G. Durrett (Oct. 2018). 'Effective Use of Context in Noisy Entity Linking'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1024–1029. (Visited on 02/03/2021) (cit. on p. 131).

Muromägi, A., K. Sirts, S. Laur (2017). 'Linear Ensembles of Word Embedding Models'. In: *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*. Linköping University Electronic Press, pp. 96–104 (cit. on pp. 34, 69).

Nebot, V., R. Berlanga (Dec. 2012). 'Exploiting Semantic Annotations for Open Information Extraction: An Experience in the Biomedical Domain'. In: *Knowledge and Information Systems* 38.2, pp. 365–389. (Visited on 06/24/2016) (cit. on p. 7).

Ng, A. Y. (2004). 'Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance'. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. New York, NY, USA: ACM, pp. 78–. (Visited on 09/25/2019) (cit. on p. 71).

Nguyen, T., M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng (2016). 'MS MARCO: A Human Generated MAchine Reading COmprehension Dataset'. In: *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-Located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*. Vol. Vol-1773. Barcelona, Spain: CEUR-WS, p. 10 (cit. on p. 63).

Niklaus, C., B. Bermeitinger, S. Handschuh, A. Freitas (Dec. 2016). 'A Sentence Simplification System for Improving Relation Extraction'. In: *Proceedings of the 26th International Conference on Computational Linguistics* 26. (Visited on 01/03/2017) (cit. on p. 25).

Niklaus, C., M. Cetto, A. Freitas, S. Handschuh (Aug. 2018). 'A Survey on Open Information Extraction'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 3866–3878. arXiv: 1806.05599 (cit. on pp. 24, 25).

Nivre, J., M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira (2016). 'Universal Dependencies v1: A Multilingual Treebank Collection'. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666 (cit. on pp. 7, 23).

Oakden-Rayner, L., G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, L. J. Palmer (May 2017). 'Precision Radiology: Predicting Longevity Using Feature Engineering and Deep Learning Methods in a Radiomics Framework'. In: *Scientific Reports* 7.1, p. 1648. (Visited on 08/25/2021) (cit. on pp. 20, 98).

Pal, H., Mausam (June 2016). 'Demonyms and Compound Relational Nouns in Nominal Open IE'. In: *5th Workshop on Automated Knowledge Base Construction (AKBC)*. San Diego, California. (Visited on 12/02/2016) (cit. on p. 25).

Pang, B., L. Lee (2004). 'A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts'. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 271 (cit. on p. 76).

– (2005). 'Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales'. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 115–124 (cit. on p. 76).

Pappu, A., R. Blanco, Y. Mehdad, A. Stent, K. Thadani (2017). 'Lightweight Multilingual Entity Extraction and Linking'. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. WSDM '17. New York, NY, USA: ACM, pp. 365–374. (Visited on 03/27/2018) (cit. on pp. xv, 63, 65, 68, 74, 143).

Patro, B. N., V. K. Kurmi, S. Kumar, V. Namboodiri (Aug. 2018). 'Learning Semantic Sentence Embeddings Using Sequential Pair-wise Discriminator'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2715–2729. (Visited on 09/24/2019) (cit. on p. 78).

Peng, Y., M. Torii, C. H. Wu, K. Vijay-Shanker (2014). 'A Generalizable NLP Framework for Fast Development of Pattern-Based Biomedical Relation Extraction Systems'. In: *BMC Bioinformatics* 15, p. 285. (Visited on 08/22/2016) (cit. on p. 58).

Peng, Y., S. Yan, Z. Lu (Aug. 2019). '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*. Florence, Italy: Association for Computational Linguistics, pp. 58–65. arXiv: 1906.05474 (cit. on pp. 32, 33, 128, 129).

Pennington, J., R. Socher, C. Manning (2014). 'Glove: Global Vectors for Word Representation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (cit. on pp. 25, 30, 67).

Perone, C. S., R. Silveira, T. S. Paula (June 2018). 'Evaluation of Sentence Embeddings in Downstream and Linguistic Probing Tasks'. In: *arXiv:1806.06259 [cs]*. arXiv: 1806.06259 [cs]. (Visited on 10/02/2018) (cit. on p. 73).

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer (Mar. 2018). 'Deep Contextualized Word Representations'. In: *Proc. of NAACL*, p. 11. arXiv: 1802.05365. (Visited on 03/26/2021) (cit. on pp. 4, 15, 26, 31, 32, 34, 63, 65–68, 74, 103, 143).

Peters, M., W. Ammar, C. Bhagavatula, R. Power (2017). 'Semi-Supervised Sequence Tagging with Bidirectional Language Models'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1756–1765. (Visited on 05/03/2019) (cit. on pp. 30, 32).

Pfeiffer, J., A. Kamath, A. Rücklé, K. Cho, I. Gurevych (2021). 'AdapterFusion: Non-Destructive Task Composition for Transfer Learning'. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 487–503. arXiv: 2005.00247. (Visited on 12/09/2022) (cit. on p. 143).

Pink, G., J. Nothman, J. R. Curran (2014). 'Analysing Recall Loss in Named Entity Slot Filling'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: ACL, pp. 820–830 (cit. on pp. 2, 24, 25, 49, 94).

Qian, Z., P. Li, Q. Zhu, G. Zhou, Z. Luo, W. Luo (2016). 'Speculation and Negation Scope Detection via Convolutional Neural Networks'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 815–825 (cit. on p. 131).

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever (June 2018). *Improving Language Understanding by Generative Pre-Training*. Technical Report. OpenAI, p. 12 (cit. on pp. 34, 63).

Radford, A., J. Wu, D. Amodei, J. Clark, M. Brundage, I. Sutskever (Feb. 2019). *Better Language Models and Their Implications*. (Visited on 04/26/2019) (cit. on p. 143).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever (Feb. 2019). *Language Models Are Unsupervised Multitask Learners*. Technical Report. OpenAI, p. 24 (cit. on pp. 5, 8, 31, 66, 67, 143).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu (2020). 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Trans-

former'. In: *Journal of Machine Learning Research* 21.140, pp. 1–67. arXiv: 1910.10683 (cit. on pp. 8, 31, 34, 143).

Rajani, N. F., M. Bornea, K. Barker (Aug. 2017). 'Stacking With Auxiliary Features for Entity Linking in the Medical Domain'. In: *BioNLP 2017*. Vancouver, Canada, Association for Computational Linguistics, pp. 39–47. (Visited on 02/03/2021) (cit. on pp. 105, 131).

Rajkomar, A., E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, J. Dean (May 2018). 'Scalable and Accurate Deep Learning with Electronic Health Records'. In: *npj Digital Medicine* 1.1, pp. 1–10. (Visited on 08/26/2021) (cit. on p. 21).

Ramshaw, L. A., M. P. Marcus (1995). 'Text Chunking Using Transformation-Based Learning'. In: *Proceedings of the 3rd Workshop on Very Large Corpora*. Vol. 3. Cambridge, Massachusetts, USA: ACL. (Visited on 11/28/2015) (cit. on p. 104).

Ratinov, L., D. Roth (2009). 'Design Challenges and Misconceptions in Named Entity Recognition'. In: *CoNLL'09*. ACL, pp. 147–155. (Visited on 07/12/2016) (cit. on p. 94).

Rettig, L., J. Audiffren, P. Cudré-Mauroux (Nov. 2019). 'Fusing Vector Space Models for Domain-Specific Applications'. In: *Proceedings of the 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. Portland, Oregon: IEEE. arXiv: 1909.02307. (Visited on 09/17/2019) (cit. on pp. 34, 69, 73, 143).

Riedel, S., L. Yao, A. McCallum, B. M. Marlin (2013). 'Relation Extraction with Matrix Factorization and Universal Schemas'. In: (visited on 05/06/2016) (cit. on p. 43).

Ronicke, S., M. C. Hirsch, E. Türk, K. Larionov, D. Tientcheu, A. D. Wagner (Mar. 2019). 'Can a Decision Support System Accelerate Rare Disease Diagnosis? Evaluating the Potential Impact of Ada DX in a Retrospective Study'. In: *Orphanet Journal of Rare Diseases* 14. (Visited on 05/27/2020) (cit. on pp. 1, 19, 21, 115, 130).

Roy, A., Y. Park, T. Lee, S. Pan (2019). 'Supervising Unsupervised Open Information Extraction Models'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 728–737. (Visited on 11/29/2021) (cit. on p. 26).

Rubenstein, H., J. B. Goodenough (Oct. 1965). 'Contextual Correlates of Synonymy'. In: *Communications of the ACM* 8.10, pp. 627–633. (Visited on 11/29/2021) (cit. on p. 27).

Saha, S. (2018). 'Open Information Extraction from Conjunctive Sentences'. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2288–2299 (cit. on p. 25).

Sahlgren, M. (2008). 'The Distributional Hypothesis'. In: *Italian Journal of Linguistics* 20.1, pp. 33–54. (Visited on 11/16/2015) (cit. on pp. 27, 28, 32).

Salem, H. A., G. Caddeo, J. McFarlane, K. Patel, L. Cochrane, D. Soria, M. Henley, J. Lund (Sept. 2018). 'A Multicentre Integration of a Computer-Led Follow-up of Prostate Cancer Is Valid and Safe'. In: *BJU international* 122.3, pp. 418–426 (cit. on pp. 1, 19, 130).

Salton, G., A. Wong, C.-S. Yang (1975). 'A Vector Space Model for Automatic Indexing'. In: *Communications of the ACM* 18.11, pp. 613–620 (cit. on pp. 28, 29).

Salton, G., C.-S. Yang, C. T. Yu (1975). 'A Theory of Term Importance in Automatic Text Analysis'. In: *Journal of the American society for Information Science* 26.1, pp. 33–44 (cit. on pp. 28, 29).

Sanh, V., A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, A. M. Rush (2022). 'Multitask Prompted Training Enables Zero-Shot Task Generalization'. In: *International Conference on Learning Representations*. Virtual. arXiv: 2110.08207. (Visited on 10/23/2021) (cit. on p. 34).

Sarmiento, R. F., F. Dernoncourt (2016). 'Improving Patient Cohort Identification Using Natural Language Processing'. In: *Secondary Analysis of Electronic Health Records*. Ed. by MIT Critical Data. Cham: Springer International Publishing, pp. 405–417. (Visited on 05/20/2020) (cit. on pp. 6, 19, 21, 108, 119).

Sarrouti, M., S. Ouatik El Alaoui (Apr. 2017). 'A Passage Retrieval Method Based on Probabilistic Information Retrieval Model and UMLS Concepts in Biomedical Question Answering'. In: *Journal of Biomedical Informatics* 68, pp. 96–103. (Visited on 08/21/2022) (cit. on p. 11).

Saussure, F. de, C. Bally, A. Sechehaye, A. Riedlinger, R. Harris (1983). *Course in general linguistics* (cit. on p. 27).

Schmidt, D., B. Osmanodja, M. Pfefferkorn, V. Graf, D. Raschke, W. Duettmann, M. G. Naik, C. J. Gethmann, M. Mayrdorfer, F. Halleck, L. Liefeldt, P. Glander, O. Staeck, M. Mallach, M. Peuker, K. Budde (Apr. 2021). 'TBase - an Integrated Electronic Health Record and Research Database for Kidney Transplant Recipients'. In: *Journal of Visualized Experiments: JoVE* 170 (cit. on pp. 3, 5, 40).

Schneider, R., S. Arnold, T. Oberhauser, T. Klatt, T. Steffek, A. Löser (2018). 'SmartMD: Neural Paragraph Retrieval of Medical Topics'. In: *Companion of the The Web Conference 2018 on The Web Conference 2018*. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 203–206 (cit. on pp. 87, 99, 111, 113, 124).

Schneider, R., C. Guder, T. Kilias, A. Löser, J. Graupmann, O. Kozachuk (2016). 'Interactive Relation Extraction in Main Memory Database Systems'. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Vol. 26. Systems Demonstrations, pp. 103–106. (Visited on 01/03/2017) (cit. on p. 37).

Schneider, R., M. Mayrdorfer, H. Schmidt, K. Budde, F. A. Gers, S. Staab (2022). 'SmartMD: Deep Learning Enabled Differential Diagnosis'. In: *To Appear*, p. 20 (cit. on p. 87).

Schneider, R., T. Oberhauser, P. Grundmann, F. A. Gers, A. Loeser, S. Staab (May 2020). 'Is Language Modeling Enough? Evaluating Effective Embedding Combinations'. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Vol. 12. Marseille, France: European Language Resources Association, pp. 4741–4750 (cit. on pp. 61, 65, 99).

Schneider, R., T. Oberhauser, T. Klatt, F. A. Gers, A. Löser (2017a). 'Analysing Errors of Open Information Extraction Systems'. In: *Building Linguistically Generalizable NLP Systems*. Copenhagen, Denmark (cit. on pp. 37, 60, 137).

– (2017b). 'RelVis: Benchmarking OpenIE Systems.' In: *International Semantic Web Conference (Posters, Demos & Industry Tracks)* (cit. on pp. 37, 60, 137).

Schumacher, E., M. Dredze (Dec. 2019). 'Learning Unsupervised Contextual Representations for Medical Synonym Discovery'. In: *JAMIA Open* 2.4, pp. 538–546. (Visited on 02/03/2021) (cit. on pp. 33, 131).

Schumacher, E., A. Mulyar, M. Dredze (July 2020). 'Clinical Concept Linking with Contextualized Neural Representations'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8585–8592. (Visited on 02/03/2021) (cit. on pp. 23, 33, 98, 105, 131).

Schütze, H., J. O. Pedersen (1995). 'Information Retrieval Based on Word Senses'. In: *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. (Cit. on p. 27).

Schwartz, R., J. Dodge, N. A. Smith, O. Etzioni (Nov. 2020). 'Green AI'. In: *Communications of the ACM* 63.12, pp. 54–63. (Visited on 11/21/2021) (cit. on p. 67).

Sennrich, R., B. Haddow, A. Birch (Aug. 2016). 'Neural Machine Translation of Rare Words with Subword Units'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for

Computational Linguistics, pp. 1715–1725. (Visited on 09/02/2021) (cit. on pp. 22, 31).

Sergeeva, E., H. Zhu, A. Tahmasebi, P. Szolovits (2019). 'Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text'. In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pp. 178–187 (cit. on p. 131).

Sharafoddini, A., J. A. Dubin, J. Lee (Mar. 2017). 'Patient Similarity in Prediction Models Based on Health Data: A Scoping Review'. In: *JMIR Medical Informatics* 5.1. (Visited on 05/22/2020) (cit. on pp. 4, 18, 19, 21, 99, 119, 132).

Sheikhshabbafghi, G., I. Birol, A. Sarkar (2018). 'In-Domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition'. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Brussels, Belgium: Association for Computational Linguistics, pp. 160–164. (Visited on 02/20/2019) (cit. on pp. 67, 68, 83).

Shen, F., H. Liu (2018). 'Incorporating Knowledge-Driven Insights into a Collaborative Filtering Model to Facilitate the Differential Diagnosis of Rare Diseases'. In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association, p. 1505 (cit. on p. 1).

Shi, H. (2017). 'Towards Automated ICD Coding Using Deep Learning'. In: *CoRR* abs/1711.04075. arXiv: 1711.04075 (cit. on pp. 21, 130).

Shickel, B., P. J. Tighe, A. Bihorac, P. Rashidi (Sept. 2018). 'Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis'. In: *IEEE journal of biomedical and health informatics* 22.5, pp. 1589–1604 (cit. on pp. 1, 2, 5, 6, 18, 130).

Si, Y., J. Wang, H. Xu, K. Roberts (Nov. 2019). 'Enhancing Clinical Concept Extraction with Contextual Embeddings'. In: *Journal of the American Medical Informatics Association* 26.11, pp. 1297–1304. (Visited on 02/16/2021) (cit. on pp. 23, 104, 131).

Sill, A. (Sept. 2016). 'The Design and Architecture of Microservices'. In: *IEEE Cloud Computing* 3.5, pp. 76–80 (cit. on p. 103).

Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan (Aug. 2023). 'Large Language Models Encode Clinical Knowledge'. In: *Nature* 620.7972, pp. 172–180. (Visited on 11/05/2023) (cit. on p. 144).

Smith, A., V. Kumar, J. Boyd-Graber, K. Seppi, L. Findlater (2018). 'Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System'. In: *23rd International Conference on Intelligent User Interfaces*. IUI '18. New York, NY, USA: ACM, pp. 293–304. (Visited on 08/30/2018) (cit. on p. 144).

Socher, R., A. P. J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts (2013). 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank'. In: *EMNLP'13*. ACL, pp. 1631–1642 (cit. on p. 75).

Soderland, S., J. Gilmer, R. Bart, O. Etzioni, D. S. Weld (2013). 'Open Information Extraction to KBP Relations in 3 Hours'. In: *Text Analysis Conference*. (Visited on 04/15/2016) (cit. on p. 25).

Soysal, E., J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu (Mar. 2018). 'CLAMP – a Toolkit for Efficiently Building Customized Clinical Natural Language Processing Pipelines'. In: *Journal of the American Medical Informatics Association* 25.3, pp. 331–336. (Visited on 07/27/2020) (cit. on p. 7).

Speer, R., J. Chin, C. Havasi (Feb. 2017). 'ConceptNet 5.5: An Open Multilingual Graph of General Knowledge'. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 4444–4451. (Visited on 11/07/2021) (cit. on p. 43).

Stanovsky, G., I. Dagan (2016). 'Creating a Large Benchmark for Open Information Extraction'. In: *EMNLP'16*. Austin, Texas: ACL, (to appear) (cit. on pp. xv, 38, 48, 51).

Stanovsky, G., I. Dagan, Mausam (2015). 'Open IE as an Intermediate Structure for Semantic Tasks'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)* (cit. on p. 25).

Stanovsky, G., J. Michael, L. Zettlemoyer, I. Dagan (2018). 'Supervised Open Information Extraction'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 885–895 (cit. on pp. 24, 26).

Starlinger, J., M. Kittner, O. Blankenstein, U. Leser (Jan. 2017). 'How to Improve Information Extraction from German Medical Records'. In: *it - Information Technology* 59.4. (Visited on 02/15/2019) (cit. on pp. 2, 4–6, 9, 60, 64, 67).

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii (Apr. 2012). 'Brat: A Web-based Tool for NLP-Assisted Text Annotation'. In: *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association for Computational Linguistics (cit. on p. 54).

Subramanian, S., A. Trischler, Y. Bengio, C. J. Pal (2018). 'Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning'. In: *International Conference on Learning Representations*. Vancouver, BC, Canada (cit. on p. 78).

Sun, J., F. Wang, J. Hu, S. Edabollahi (Dec. 2012). 'Supervised Patient Similarity Measure of Heterogeneous Patient Records'. In: *ACM SIGKDD Explorations Newsletter* 14.1, pp. 16–24. (Visited on 05/22/2020) (cit. on pp. 20, 100, 108).

Sun, M., X. Li, X. Wang, M. Fan, Y. Feng, P. Li (Apr. 2019). 'Logician: A Unified End-to-End Neural Approach for Open-Domain Information Extraction'. In: *arXiv:1904.12535 [cs]*. arXiv: 1904.12535 [cs]. (Visited on 11/29/2021) (cit. on p. 26).

Sutton, R. T., D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker (Dec. 2020). 'An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success'. In: *npj Digital Medicine* 3.1, p. 17. (Visited on 04/22/2021) (cit. on pp. 3, 5, 6, 17–19, 98, 111, 130).

Sutton, R. S., A. G. Barto (Sept. 2016). *Reinforcement Learning: An Introduction*. 1nd Edition edition. Cambridge, Mass.: A Bradford Book (cit. on p. 99).

Tang, B., H. Cao, Y. Wu, M. Jiang, H. Xu (Apr. 2013). 'Recognizing Clinical Entities in Hospital Discharge Summaries Using Structural Support Vector Machines with Word Representation Features'. In: *BMC Medical Informatics and Decision Making* 13.1, S1. (Visited on 02/24/2021) (cit. on p. 104).

Tang, S., V. R. de Sa (Dec. 2018). 'Improving Sentence Representations with Multi-view Frameworks'. In: *IRASL Colocated at NeurIPS2018*. Montréal, Canada, p. 13 (cit. on pp. 78, 81).

Triplitt, C. (Oct. 2006). 'Drug Interactions of Medications Commonly Used in Diabetes'. In: *Diabetes Spectrum* 19.4, pp. 202–211. (Visited on 09/01/2021) (cit. on pp. 7, 9).

Tseytlin, E., K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, R. S. Jacobson (2016). 'NOBLE–Flexible Concept Recognition for Large-Scale Biomedical Natural Language Processing'. In: *BMC bioinformatics* 17.1, pp. 1–15 (cit. on pp. 6, 105, 131).

Uzuner, Ö., B. R. South, S. Shen, S. L. DuVall (2011). '2010 I2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text'. In: *Journal of the American Medical Informatics Association : JAMIA* 18.5, pp. 552–556. (Visited on 02/18/2021) (cit. on p. 131).

van Aken, B., J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. Gers, A. Loeser (Apr. 2021). 'Clinical Outcome Prediction from Admission Notes Using Self-Supervised Knowledge Integration'. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 881–893. (Visited on 04/22/2021) (cit. on pp. 19, 21, 98, 99, 107, 120, 130).

van Aken, B., I. Trajanovska, A. Siu, M. Mayrdorfer, K. Budde, A. Loeser (June 2021). 'Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?' In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. Online: Association for Computational Linguistics, pp. 35–40 (cit. on pp. 5, 7, 100, 107, 131).

van Aken, B., B. Winter, A. Löser, F. A. Gers (2019). 'How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations'. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. New York, NY, USA: ACM, pp. 1823–1832. (Visited on 11/20/2019) (cit. on pp. 30, 72).

van der Vegt, A., G. Zuccon, B. Koopman, A. Deacon (Oct. 2020). 'How Searching under Time Pressure Impacts Clinical Decision Making'. In: *Journal of the Medical Library Association : JMLA* 108.4, pp. 564–573. (Visited on 08/07/2022) (cit. on p. 3).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin (June 2017). 'Attention Is All You Need'. In: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA.: Curran Associates, Inc. arXiv: 1706.03762. (Visited on 09/17/2018) (cit. on pp. 5, 32, 74, 100, 107, 132).

Vibert, N., C. Ros, L. L. Bigot, M. Ramond, J. Gatefin, J.-F. Rouet (2009). 'Effects of Domain Knowledge on Reference Search with the PubMed Database: An Experimental Study'. In: *Journal of the American Society for Information Science and Technology* 60.7, pp. 1423–1447 (cit. on p. 11).

Voorhees, E. M., D. M. Tice (2000). 'Building a Question Answering Test Collection'. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 200–207 (cit. on p. 76).

Vrandečić, D., M. Krötzsch (2014). 'Wikidata: A Free Collaborative Knowledgebase'. In: *Communications of the ACM* 57.10, pp. 78–85 (cit. on p. 104).

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman (Dec. 2019). 'SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems'. In: *Proceedings of the Annual Conference on Neural Information Processing Systems 2019 {NeurIPS 2019}*. Vancouver, BC, Canada, pp. 3261–3275. arXiv: 1905.00537. (Visited on 02/23/2021) (cit. on pp. 32, 33, 98, 103, 129).

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman (2018). 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. (Visited on 11/15/2019) (cit. on pp. 32, 33, 103, 129).

Wang, L., S. Li, Y. Lv, H. Wang (2017). 'Learning to Rank Semantic Coherence for Topic Segmentation'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1340–1344. (Visited on 03/01/2018) (cit. on pp. 6, 69).

Wang, L., Y. Li, S. Lazebnik (June 2016). 'Learning Deep Structure-Preserving Image-Text Embeddings'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5005–5013. (Visited on 06/12/2018) (cit. on p. 144).

Wang, X., Y. Zhang, Q. Li, Y. Chen, J. Han (Aug. 2018). 'Open Information Extraction with Meta-pattern Discovery in Biomedical Literature'. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. Washington DC USA: ACM, pp. 291–300. (Visited on 08/16/2021) (cit. on p. 25).

Watford, S. M., R. G. Grashow, V. Y. De La Rosa, R. A. Rudel, K. P. Friedman, M. T. Martin (Aug. 2018). 'Novel Application of Normalized Pointwise Mutual Information (NPMI) to Mine Biomedical Literature for Gene Sets Associated with Disease: Use Case in Breast Carcinogenesis'. In: *Computational toxicology (Amsterdam, Netherlands)* 7, pp. 46–57. (Visited on 05/19/2021) (cit. on pp. 111, 113, 114).

Webster, J. J., C. Kit (Aug. 1992). 'Tokenization as the Initial Phase in NLP'. In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*. COLING '92. USA: Association for Computational Linguistics, pp. 1106–1110. (Visited on 02/26/2021) (cit. on p. 105).

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le (2022). 'Finetuned Language Models Are Zero-Shot Learners'. In: *International Conference on Learning Representations*. Virtual. arXiv: 2109.01652. (Visited on 10/23/2021) (cit. on p. 34).

Wei, Q., Z. Ji, Y. Si, J. Du, J. Wang, F. Tiryaki, S. Wu, C. Tao, K. Roberts, H. Xu (Mar. 2020). 'Relation Extraction from Clinical Narratives Using Pre-trained Language Models'. In: *AMIA Annual Symposium Proceedings* 2019, pp. 1236–1245. (Visited on 08/16/2021) (cit. on pp. 4, 21).

Werbos, P. J. (1990). 'Backpropagation Through Time: What It Does And How To Do It'. In: *Proceedings of the IEEE* 78.10, pp. 1550–1560 (cit. on p. 94).

Weston, J., A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, T. Mikolov (May 2016). 'Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks'. In: *Proceedings of the 4th International Conference on Learning Representations, {ICLR} 2016*. San Juan, Puerto Rico. arXiv: 1502.05698. (Visited on 11/20/2019) (cit. on p. 73).

White, A. S., D. Reisinger, K. Sakaguchi, T. Vieira, S. Zhang, R. Rudinger, K. Rawlins, B. V. Durme (2016). 'Universal Decompositional Semantics on Universal Dependencies'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (cit. on p. 47).

White, R. W., E. Horvitz (2014). 'From Health Search to Healthcare: Explorations of Intention and Utilization via Query Logs and User Surveys J Am Med Inform Assoc. 2014 January; 21 (1): 49–55'. In: *Journal of the American Medical Informatics Association: Jamia* 21.1, pp. 49–55 (cit. on p. 89).

Wiebe, J., T. Wilson, C. Cardie (2005). 'Annotating Expressions of Opinions and Emotions in Language'. In: *Language resources and evaluation* 39.2-3, pp. 165–210 (cit. on p. 75).

Wolfe, T., M. Dredze, B. Van Durme (July 2017). 'Pocket Knowledge Base Population'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 305–310. (Visited on 11/29/2021) (cit. on p. 25).

Wu, F., D. S. Weld (2010). 'Open Information Extraction Using Wikipedia'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 118–127. (Visited on 06/02/2016) (cit. on p. 25).

Wu, L., A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston (Feb. 2018). 'StarSpace: Embed All The Things!' In: *Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. New Orleans: {AAAI} Press, p. 9 (cit. on p. 34).

Wu, Y., M. Jiang, J. Xu, D. Zhi, H. Xu (Apr. 2018). 'Clinical Named Entity Recognition Using Deep Learning Models'. In: *AMIA Annual Symposium Proceedings* 2017, pp. 1812–1819. (Visited on 02/16/2021) (cit. on pp. 23, 104, 131).

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean (Oct. 2016). 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'. In: *arXiv:1609.08144 [cs]*. arXiv: `1609.08144 [cs]`. (Visited on 02/24/2021) (cit. on pp. 22, 31, 104).

Xu, Y., M.-Y. Kim, K. Quinn, R. Goebel, D. Barbosa (2013). 'Open Information Extraction with Tree Kernels.' In: *HLT-NAACL*, pp. 868–877. (Visited on 05/04/2017) (cit. on pp. xv, 48).

Yan, Z., D. Tang, N. Duan, S. Liu, W. Wang, D. Jiang, M. Zhou, Z. Li (2018). 'Assertion-Based QA with Question-Aware Open Information Extraction'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32 (cit. on p. 24).

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le (Dec. 2019). 'XLNet: Generalized Autoregressive Pretraining for Language Understanding'. In: *Proceedings of the Annual Conference on Neural Information Processing Systems 2019 {NeurIPS 2019}*. Vancouver, BC, Canada, pp. 5754–5764. arXiv: 1906.08237. (Visited on 09/24/2019) (cit. on p. 78).

Yin, W., H. Schütze (Aug. 2015). 'Learning Meta-Embeddings by Using Ensembles of Embedding Sets'. In: *arXiv:1508.04257 [cs]*. arXiv: 1508.04257 [cs]. (Visited on 02/20/2019) (cit. on pp. 34, 69, 73).

Yoo, I., A. S. M. Mosa (July 2015). 'Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users'. In: *JMIR Medical Informatics* 3.3. (Visited on 02/23/2018) (cit. on pp. 11, 89).

Yu, D., L. Huang, H. Ji (2017). 'Open Relation Extraction and Grounding'. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 854–864 (cit. on p. 25).

Yuksel, S. E., J. N. Wilson, P. D. Gader (Aug. 2012). 'Twenty Years of Mixture of Experts'. In: *IEEE Transactions on Neural Networks and Learning Systems* 23.8, pp. 1177–1193 (cit. on p. 131).

Zhan, J., H. Zhao (2020). 'Span Model for Open Information Extraction on Accurate Corpus'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 9523–9530 (cit. on p. 26).

Zhang, N., G. Yang, Z. Gao, C. Xu, Y. Zhang, R. Shi, J. Keegan, L. Xu, H. Zhang, Z. Fan, D. Firmin (June 2019). 'Deep Learning for Diagnosis of Chronic Myocardial Infarction on Nonenhanced Cardiac Cine MRI'. In: *Radiology* 291.3, pp. 606–617. (Visited on 11/29/2021) (cit. on p. 20).

Zhang, Y., Q. Chen, Z. Yang, H. Lin, Z. Lu (May 2019). 'BioWordVec, Improving Biomedical Word Embeddings with Subword Information and MeSH'. In: *Scientific Data* 6.1, p. 52. (Visited on 10/09/2021) (cit. on p. 31).

Zhao, H., Z. Lu, P. Poupart (July 2015). 'Self-Adaptive Hierarchical Sentence Model'. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence {IJCAI 2015}*. Buenos Aires, Argentina, pp. 4069–4076. arXiv: 1504.05070 (cit. on pp. 78, 81).

Zhu, D., S. Wu, B. Carterette, H. Liu (June 2014). 'Using Large Clinical Corpora for Query Expansion in Text-Based Cohort Identification'. In: *Journal of Biomedical Informatics* 49, pp. 275–281. (Visited on 05/19/2020) (cit. on p. 21).

List of Figures

List of Tables

Appendix A

Related Contributions

We explored applications of our methods, recreated experiments from the literature, and created open-source implementations as additional practical contributions to the central part of this thesis.

A.1. Open Source Contributions

- TeXoo: `https://github.com/sebastianarnold/TeXoo`
- TeXooPy: `https://github.com/DATEXIS/TeXooPy`
- PubMedSection: `https://github.com/DATEXIS/pubmedsection`
- UMLS Parser: `https://github.com/DATEXIS/UMLSParser`
- SentEval k8s: `https://github.com/DATEXIS/SentEval-k8s`
- OpenNLP: `https://github.com/apache/opennlp/pull/337`
- SmartMD DDx: `https://github.com/DATEXIS/smartmd-backend`
- Discovered and assisted resolving multiple bugs in DeepLearning4J
  - `https://github.com/eclipse/deeplearning4j/issues/7125`
  - `https://github.com/eclipse/deeplearning4j/issues/7120`
  - `https://github.com/eclipse/deeplearning4j/issues/7001`

A.2. Supervised Theses

We supervised the following selected bachelor's and master's theses in context to this thesis. They cover applications of our methods, provide additional experiments and explore related work by recreating experiments.

- T. Klatt, "Benchmarking Transformers for Biomedical Text Understanding", Masters Thesis, Beuth University of Applied Sciences, Berlin, Germany, 2022.
- T. Steffek, "Neural Facet Detection on Medical Resources," Bachelor Thesis, Beuth University of Applied Sciences, Berlin, Germany, 2019.
- T. Oberhauser, "Neural Information Retrieval with Vector Space Queries," Master Thesis, Beuth University of Applied Sciences, Berlin, Germany, 2019
- T. Schilling, "Information Retrieval mit Satz- und Wortembeddings," Master Thesis, Beuth University of Applied Sciences, Berlin, Germany, 2018.
- T. Klatt, "A Graphical Training Interface for Named Entity Recognition," Bachelor Thesis, Beuth University of Applied Sciences, Berlin, Germany, 2018.