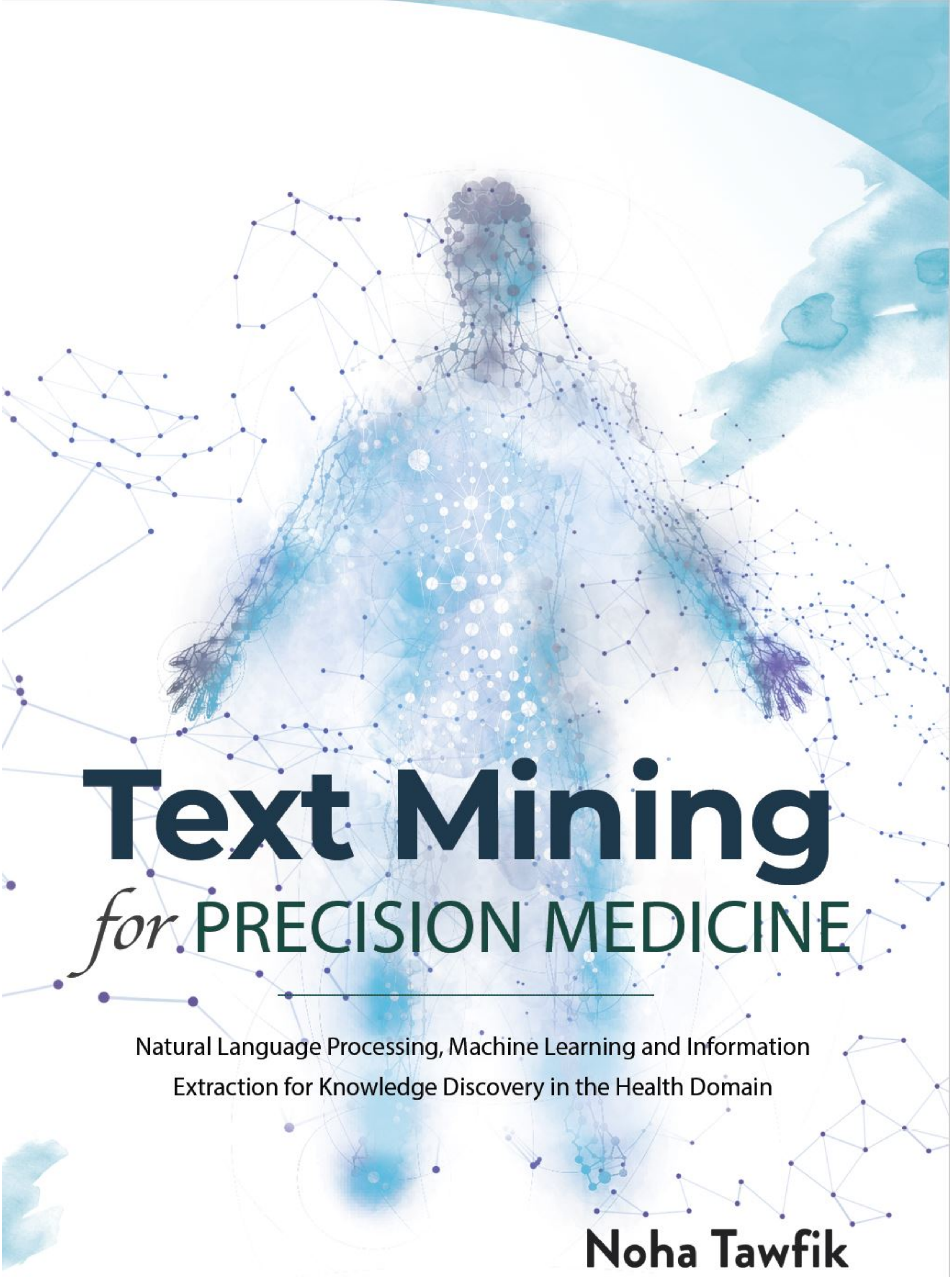


Text Mining *for* PRECISION MEDICINE

Natural Language Processing, Machine Learning and Information Extraction for Knowledge Discovery in the Health Domain

Noha Tawfik

Text Mining for Precision Medicine

Noha Seddik Tawfik

Cover design: Yerai Ibarria (https://www.fiverr.com/yeraiibarria)

Text Mining for Precision Medicine
ISBN: 978-90-393-7342-2

Text Mining for Precision Medicine

Natural Language Processing, Machine Learning and Information Extraction for Knowledge Discovery in the Health Domain

Tekstanalyse voor zorgbehandelingen op maat

Natuurlijke taalverwerking, machinaal leren en informatie-extractie voor kennisontdekking in de gezondheidszorg

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 24 november 2020 des middags te 4.15 uur

door

Noha Seddik Abdelsalam Tawfik Abdelrahman

geboren op 5 november 1988
te Alexandria, Egypte

Promotor: Prof. dr. S. Brinkkemper
Copromotor: Dr. M.R. Spruit

Acknowledgments

Four years ago, carried by my passion for research, I embarked on my doctoral journey with the utmost determination to keep going for as long as it takes to succeed. Little did I know that this was not always so easy. Sometimes it felt like a mountain climb filled with struggles, hardships, and frustration. Every milestone reached was a joy on its own, but today, I find myself at the top with an utter sense of fulfillment and satisfaction. Not only am I proud for starting this journey, but reaching the mountain peak was only possible because of many people to whom I would like to express my sincere gratitude.

First and foremost, I would like to thank my supervisors ***Professor Sjaak Brinkkemper*** and ***Dr. Marco Spruit***. It has been a privilege to work with both of you. Thank you for believing in me and for giving me the chance to start this journey. ***Sjaak***, I still remember your welcoming chat in my first visit to the UU and how it made me much more comfortable and at ease. Thank you for all the nice talks we had, and for your support throughout the years, and especially in the past year. ***Marco***, I am incredibly grateful for your responsiveness that was the main reason that brought me to the Netherlands. I know how risky it was to accept me as your PhD student while I am at the other end of the world and still offer me your mentorship remotely. Thank you for inspiring me and guiding me through this unpredictable PhD roller-coaster, for the support and patience in all times, and for your confidence in letting me decide my research directions. I sure hope that I made the risk worth taking.

I am grateful to have had the privilege of studying in the prestigious University of Utrecht, which allowed me to meet some of the best and brightest individuals. I am indebted to many of my current and former colleagues at the Applied Data Science lab, ***Armel, Bilge, Chaim, Ian, Injy, Lamia, Shaheen, Vincent, and Wienand***. Thank you for the encouraging chats, lovely lunches, and for hosting me in your offices during my visits.

I am also tremendously fortunate for having a lovely and caring family, as well as supporting friends, who have helped me get through the past few years successfully. I am truly grateful for my family for their continous love, encouragement, and prayers they have sent my way along this journey. I hope I have made you proud. Also, thank you for all the help with the girls when I needed it the most.

I owe a huge thanks to a very special person, my husband ***Sherif***, for his ongoing and unfailing love and understanding. You supported me in every way possible and always endured my post-rejection rants calmly. Without you, I would have never managed to finish my degree.

Finally, ***Lara & Hana***, our beautiful twin daughters who joined us half-through my PhD journey. While kids do not make things easier, you have made me stronger, better and more fulfilled than I could have ever imagined. If you read this when you are older, I love you to the moon and back, and I will make sure you get to follow your dreams just as I did.

– Noha, September 2020

Contents

1 Introduction

Research in biomedical informatics aims at enriching biomedical science and improving the health of individuals and their quality of care. For that purpose, researchers design, implement, and evaluate various informatics methods that serve different biomedical scientific fields such as decision support systems, clinical research, human interface design, data analysis, and biomedical database management. These methods and solutions are employed to support the activities of health professionals but are also intended to improve the overall healthcare provided to the patient.

In today's world, data is the real driving force within the medical informatics domain. It is expected that the need for informatics experts who understand the collection, analysis, and manipulation of data will continue to grow (Fridsma, 2016). Two sources of data are highly relevant to the medical domain, namely the patient records and scientific literature. The rate of publication of biomedical and health sciences literature is increasing exponentially. As an indication, it is estimated that the number of clinical trials increased from 10 per day in 1975, to 55 in 1995, to 95 in 2015. In 2017, the PubMed repository contained around 27 million articles, 2 million medical reviews, 500,000 clinical trials, and 70,000 systematic reviews (Catillon, 2017a). Subsequently, this leads to the development of web-based search tools such as PubMed and Google Scholar (Singhal, Simmons, & Lu, 2016). Similarly, electronic health records adoption rates, by hospitals and health institutes, increased dramatically in the last decade. The global widespread digitization of EHRs enabled the use of data mining tools to analyze patient records for valuable knowledge (Ross, Wei, & Ohno-Machado, 2014). Moreover, by providing access to medical records warehouses through internet clouds, data can be collected and analyzed much more quickly on a wider range. Consequently, it could be possible to apply preventative measures or even be possible to predict epidemics and act more quickly in the future (Simmons, Singhal, & Lu, 2016).

Generally, one can refer to all informatics research within the medical domain as biomedical informatics. The American Medical Informatics Association (AMIA) categorizes the methods and techniques developed by biomedical informatics scientists into five major areas of research according to the application sub-domain. These categories are Translational Bioinformatics, Clinical Research Informatics, Clinical Informatics, Consumer Health Informatics, Public Health Informatics (*The Science of Informatics — AMIA*, 2019) This work falls within the second category: the clinical research informatics that promotes medical knowledge discovery through the use of informatics. It spans a variety of applications, including information management of data related to clinical trials, information extraction from medical resources, and informatics activities to support translational research (*Clinical Research Informatics — AMIA*, 2019). Our research in biomedical informatics follows a solution-driven approach that builds on and contributes to existing information sciences and technologies, emphasizing their application in the medical domain.

1.1 Medical Research Domain: Precision Medicine

Since the late 1990s, Evidence-Based Medicine (EBM) has been the dominant practice for clinical and decision-making processes in health care. EBM is defined as the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" rather than making clinical decisions solely on personnel clinical experience (Stelter & Stelter, 2014). Developments in genomics and other molecular sciences, as well as new sequencing technologies and individual data measurements, led to the promotion of Precision Medicine (PM). The new paradigm includes significant changes in therapeutics: from one-size-fits-all that only benefits the "average" patient to "targeted treatments" based on individual patient characteristics (Douglas Zhang, 2015). Consequently, a shift in the clinical treatment process follows from a trial-and-error approach to "the right treatment, the right patient, at the right time" concept. The PM approach complements EBM rather than opposes it; by collecting and using more individual data for therapy, diagnosis, and prognosis, it places the individual more clearly at the center of the production of the knowledge of medicine (Chow, Gallo, & Busse, 2018). Adopting PM into practice leads to the cost-benefit optimization of treatment or prevention of diseases. Moreover, It has significant effects as it influences economy, pharmaceutical industry, health institutions, and political and social agendas. This was evident by the number of international initiatives such as the British "100000 GENOME" project, announced in 2012, followed by the US "Precision Medicine" program initiated by its former president Barack Obama in 2015. In the same year, the French plan "France genomic medicine 2025" was also announced with the aims of positioning France, within ten years, in the leading pack of countries engaged in genomic medicine (Vamathevan & Birney, 2017). The terms "predictive medicine," "personalized medicine," "individualized medicine," or "precision medicine" are often used interchangeably in the medical literature to characterize this targeted-approach of medicine. Implicitly, all these definitions aim at improving the predictive aspects and the clinical evolution by integrating the therapeutic responses and the potential side effects of the different treatments (Schleidgen, Klingler, Bertram, Rogowski, & Marckmann, 2013). Throughout this work, the term precision medicine is used, as it conveys a more precise description of the new trend, namely grouping genotypes and phenotypes into subgroups according to available data.

This research focuses on analyzing data that supports the precision medicine (PM) concept. Better knowledge of population genetics, health, environment, and personal lifestyle allows for more targeted care (Chawla & Davis, 2013). Subsequently, the PM research generate findings, facts, and results that are also added to the massive pool of data. Today, five years after the US introduced the Precision Medicine Initiative, a Google search with the term "precision medicine" returns 52 million pages, and on the PubMed site, the same term identifies 25,000 articles indexed in the scientific articles.

1.2 Computing Research Domain: Text Mining

In addition to the health research domain, this dissertation employs and contributes to methods and techniques from various research domains within the Text Mining (TM) field. Text mining helps in the acquisition and extraction of hidden information or trends that researchers fail to capture from large-scale and content-rich text data (Aggarwal & Zhai, 2012). Assigning a comprehensive definition to TM is hard since the field emerged out of multiple related but distinct disciplines including data mining techniques, information extraction, information retrieval, machine learning, natural language processing, computational linguistics, statistical data analysis, linear geometry, probability theory, and even Graph Theory(Miner, Elder, & Nisbet, 2012). It could also be referred to as text analytics, intelligent text analysis, text data mining, or knowledge Discovery in Text (KDT), but they all refer to the process of analyzing and processing semi-structured and unstructured text data. Figure 1.1 illustrates the major fields that overlap with text mining (Miner et al., 2012). This section introduces in-depth three of the related field relevant to the research conducted in this thesis: Natural Language Processing (NLP), Information Extraction (IE) and Machine Learning (ML).

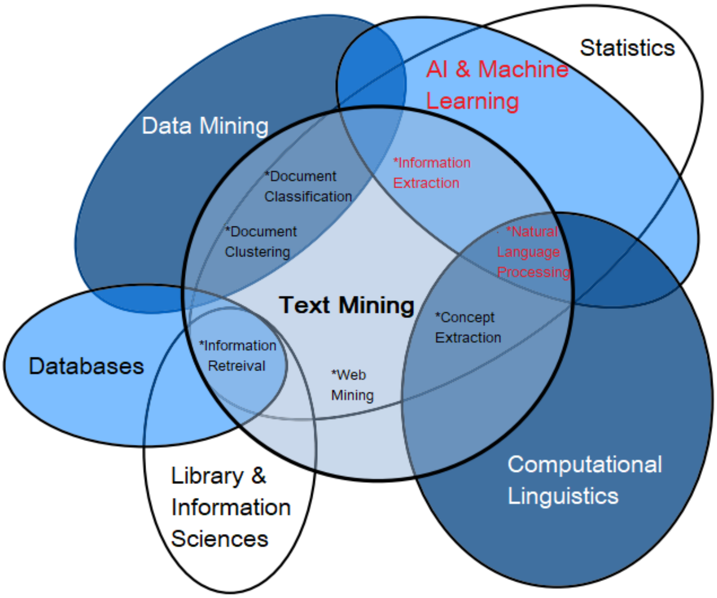


Figure 1.1: The overlap of seven fields related to text mining. The scientific fields in which this research is conducted are highlighted in red.

1.2.1 Natural Language Processing

Under the umbrella of automatic natural language processing (NLP), all the research and development aims at modeling and reproducing, with the help of machines, the human capacity to produce and understand linguistic utterances for communication purposes. Historically, the first attempt in the field of NLP, in 1954, has focused on machine translation, with the development of a naive automatic translator between the Russian and English langauges. Although the vocabulary included only 250 words, and the grammar had 6 rules, this experiment triggered many works in the domain. The 1970s witnessed the development of semantic and syntactic approaches with the emergence of grammatical formalisms (Nadkarni, Ohno-Machado, & Chapman, 2011). Today, the field of NLP is in constant change with substantial breakthroughs such as the pre-trained deep learning solutions. From the past to the future, NLP has steadily been ,and still is, delivering its high promises in making data more user-friendly and conversational. Subsequently, more and more mainstream users will adopt NLP-driven techniques in daily activities (Kamath, Liu, & Whitaker, 2019).

In natural language, the facts expressed in the text are not necessarily all explicit: the human reader infers the missing elements given his common sense of knowledge and experience. An example is textual inference, defined as the relationship between two fragments of text modeled as a three-class prediction problem. Given two snippets of text: premise and hypothesis, textual inference determines whether the hypothesis is true (*Entailment*), false (*contradiction*), or undetermined (*Neutral*) with respect to the premise (Dagan, Roth, Sammons, & Zanzotto, 2013a). Traditional NLP techniques do not naturally have the capability of detecting inference. Natural language inference (NLI), also known as recognizing textual entailment (RTE) require a deep understanding of the semantic similarity between the hypothesis and the premise. NLI models aim to analyze lexical, syntactic, and semantic features of the text independent of the given application, leading to a better understanding of the linguistic variability of the language(Norvig, 1987).

It is essential to differentiate between the general language, that we speak every day, and the specialized languages specific to each sector, such as technical, scientific, legal, financial, and medical language. The scientific and medical languages have specificities, both in terms of style and vocabulary, that should be taken into account when processing medical text. The style depends on the text source; it will be rather informative and impersonal if the source is patient records or clinical notes, while it will be more complex and well-argued if it is an essay or an article published in a scientific or medical journal intended to be read by specialists. In the second case, it will show the opinion of the author(s) in order to convince the readers with a hypothesis supported by arguments and examples (Wu & Liu, 2011). On the other hand, patient authored text, i.e. posts of medical forums or tweets tend to be very similar to the general language structure and contain words of everyday language, which patients use in their particular sense. Biomedical natural language processing or BioNLP, refers to the methods and study of how text mining may be applied to texts and literature of the biomedical and clinical domains (K. B. Cohen & Demner-Fushman, 2014). As highlighted above, the processing of medical language is complex because it contains a significant number of abbreviations, synonyms, acronyms, ambiguous morphology, and rich semantics.

Throughout this dissertation, many existing NLP methods are applied and, where possible, new frameworks are constructed for the understanding of English medical text. In Chapter 2 for example, we design and implement a rule-based NLP method to extract genetic information from scientific text automatically while chapter 6 compares existing sentence embeddings in capturing biomedical knowledge from text. One of the main focus points of this work was to investigate and contribute to the medical language inference domain. Research in medical NLI is at the heart of several applications of BioNLP such as biomedical information extraction, automatic summary of scientific literature, clinical question answering, and text paraphrasing (Dagan, Dolan, Magnini, & Roth, 2009). Chapters 5 and 7 employ state-of-the-art text representation techniques and deep networks to improve the detection of inference among pairs of medical and clinical text.

1.2.2 Information Extraction

Information Extraction (IE) techniques have matured considerably in the last decade. They involve extracting precise information from documents and structuring them into a predefined form (Pazienza, 1999). This usually involves deducing characteristics concerning entities or events mentioned in the texts and the relations between them. In the wide range of automatic textual document processing, it will be convenient to situate IE, both in terms of its objectives and methods, as an intermediate level between document retrieval on the one hand, and automatic comprehension, on the other hand (Gaizauskas & Wilks, 1998). In document retrieval, the general objective is to facilitate the selection of a subset of relevant documents in a database in response to a query from the users. The output is the document itself, without understanding or interpreting its content. In contrast, in automatic comprehension the goal is to obtain a representation of the meaning of the given text. This objective calls for an exhaustive analysis, syntactic and semantic, of each sentence and the relations that they maintain, and, in some cases, the construction of a knowledge base. In IE, it is also required to make sense out of documents but applied to targeted parts for particular tasks, not the text as a whole (Cowie & Wilks, 2000). IE thus represents a good compromise, likely leading to improvements in content analysis of textual documents, complementary to the document retrieval. The IE field has grown significantly; a large scientific community has thus been formed, laying solid methodological foundations and allowing an accumulation of techniques and software systems.

In the medical domain, it is desirable to harvest information and knowledge from unstructured data to support automated systems and to build decision support systems. The importance of information extraction methods in the medical domain is evident with the increased availability of clinical and medical textual data. They follow either a rule-based approach, a machine learning-based approach, or combine both for better performance (Ghoulam, Barigou,

& Belalem, 2015; Y. Wang, Wang, et al., 2018). On the precision medicine level, many shared tasks and challenges such as the TREC and BioCreative precision medicine tracks were initiated in 2017 to promote interdisciplinary collaboration between different fields (Islamaj Dogan et al., 2017; Roberts et al., 2018). In chapters 2 and 5 of this dissertation, we use IE methodologies to extract specific information from biomedical abstracts available in the PubMed repository.

1.2.3 Machine Learning

Machine learning is a core discipline within artificial intelligence (AI) whose objective is the elaboration of automated methods that the machine will use to learn from data (Oquendo et al., 2012). Back in the late 1950s, Arthur Samuel, a researcher in computer science, proposed a definition of ML as being "a field of study that gives computers the ability to learn without having been explicitly programmed" (Wiederhold & McCarthy, 2010). In simpler words, it is rather a program that is developed to learn from its own experience. With active research, continuous development and promised solutions, ML methods are now used in a variety of fields that affect our daily lives. These methods are inspired by biological systems, among others, the neural network of our brain. Both ML and AI, although highly publicized today, are not new: their first use dates back to the Second World War and coining of the "electronic brain" concept. While the theories of deep learning were already first introduced in the 1990s, the availability of computational power and increasingly large datasets have made applying them feasible only very recently.

ML is at the heart of the medicine of the future with hopes of improving the quality of care. The rise of machine learning within the domain has been made possible by the increasing digitization of health data, computerized medical records, prescription software, medical imaging, and genetic sequencing (Deo, 2015). In this work, we apply supervised traditional machine learning algorithms in chapters 4 and 6, while we explore deep learning in chapter 6 and 7.

1.3 Research Questions

As previously mentioned, healthcare is rapidly shifting towards personalizing the medical practice. This has resulted in funds, grants, and interdisciplinary projects to accelerate the PM research. However, the PM paradigm is still in its infancy, research in the domain is scarce and in need of computational knowledge to reach its full potential. This work focuses on bridging the gap between the medical community and the BioNLP research in general and in the PM domain in particular. There is a lack of scientific methods that address the several challenges surrounding the analysis of medical textual data. Throughout the dissertation, we investigate the benefits of employing text mining and machine learning for the sake of speeding-up the knowledge discovery process in the PM domain. This research is a step forward towards adopting Precision Medicine into practice by focusing on two main goals: supporting the PM applicability and proving the efficacy of the paradigm. We, therefore, pose the following main research question:

MRQ — How can Biomedical NLP techniques support and advance the precision medicine approach through collection and analysis of clinical and medical textual resources?

To answer the main research question, we pose seven research questions, as explained below. Under the conviction that medical research should ultimately aim to improve care for the individual patient, the goal of this research is twofold. Firstly, we aim to contribute to the PM domain by obtaining valuable knowledge from unstructured resources. Secondly, we aim to apply state of the art NLP techniques to multiple data sources in order to better support the PM concept. This can ultimately lead to a better understanding of the medical data in general and how to interpret it in favor of PM in particular. The individual chapters of this dissertation all intend to contribute to these goals, with the first two questions addressing the first goal and the other five questions addressing the second goal.

RQ1 — How can text mining techniques be employed to extract relevant information from genome-wide associations studies?

Seventeen years ago, the first sequencing of the human genome was conducted in 2003 by the Human Genome Project. It had consumed a budget of $2.7 billion over thirteen years and required the collaboration of twenty research centers. Since then, associated times and costs for sequencing genomes reduced to $10,000 in 2007, then to $1,000 in 2014, and more reductions are expected (Thermes, 2014). The decline is so drastic that some compare the progress rate in genomics to that of Moore's Law in processor development. Since then, the entire landscape of genome sequencing has been revolutionized. Consequently, this led to an increase in genetic-based research and, more specifically, the rate of conducting Genome-Wide-Association-Studies (GWAS). A GWAS is a case-control study in a population of unrelated subjects. The goal is to identify Single Nucleotide Polymorphisms (SNPs), whose risk is significantly different between cases and controls. Not only would it help in understanding the disease etiologies, but GWAS results could also be used in prevention stages as it predicts personal susceptibility to disease and drug responsiveness (Ikegawa, 2012). The discovery of a large number of SNPs encouraged the construction of several databases to store the info. Nonetheless, these databases still rely on manual curation of published literature to update their content. This research investigates how to effectively apply and combine existing NLP techniques in an Information Extraction framework to help domain experts in exploring GWAS in order to obtain exploratory results needed for their future research.

RQ2 — How to structure knowledge extracted from large collections of scientific literature?

With the recent hype around precision medicine, at both the academic and public levels, debates on the correct interpretation of the domain has risen. PM suffers from conceptual vagueness and lacks a shared understanding of its knowledge among stakeholders. Such limitations complicate and delay the potential benefits of PM (Schleidgen et al., 2013). The related data sources available on the Web today are multiple and heterogeneous. Optimal use of these resources requires both computer and biological skills from users due to the lack of documentation and difficulties in interacting with data sources (Schuurman & Leszczynski, 2008). This process involves modeling and formalizing domain knowledge into an ontology. Ontologies are data models that transform a domain's data into machine-readable representations to describe how a domain's information is organized. In an attempt to generate a standard reference for the use of the PM concepts and vocabulary, this research employs web technology semantics to build an ontology for precision medicine.

RQ3 — How to incorporate text mining methods in detecting contradictory statements found in scientific literature?

Investigating the same scientific experiment, from different sources, in the medical field, i.e., medications responses, interventions, or adverse drug reactions for a specific case study, not all outcomes are identical (Prasad, Cifu, & Ioannidis, 2012b). This research hypothesizes that highlighting different outcomes to the same medical practice supports the PM claims that there is no ”one-size-fits-all” treatment strategy. However, the manual curation of evidence-based medicine from literature is exhausting, costly, and time-consuming, even in a specific sub-domain. For this reason, automatic analysis of the medical text with NLP tools is a common approach in order to speed up the process (Collier, Nazarenko, Baud, & Ruch, 2006). This research question aims to investigate to what extent machine learning models can automatically identify conflicting statements among published evidence-based medicine findings.

RQ4 — To what extent can deep learning improve the detection of textual inference, compared to traditional machine learning techniques?

Detecting medical inference is of great value to both the open and medical domains and specifically to precision medicine. This part extends the prior question RQ3 to model entailment and neutral relations along with contradiction inference between sentences. This research compares and evaluates different machine and deep learning techniques in detecting inference among pairs of medical sentences. Moreover, it investigates whether combining traditional features with deep networks have an additional benefit over classical machine learning techniques and deep learning techniques when applied to predicting inference in the biomedical text.

RQ5 — Which textual representations are most suited for the biomedical domain?

The complexity of the medical language comes from the rich vocabulary and the conceptual overlap among its terms, but it is above all functional. It allows for effective communication as health professionals can condense a large amount of information in a short text (Chaussabel, 2004). The importance of text representations in biomedical language processing becomes evident, given the amount of available data. Over the last two years, the NLP community has witnessed a ”revolution” of text embeddings with the availability of context-dependent representations. Most NLP tasks have experienced an unusual performance jump, with an emerging shift for sentence /paragraph processing. This research question, therefore, assesses the ability of a range of sentence representations models to capture the rich and complex semantics of medical sentences.

RQ6 — What are the benefits of using state-of-the-art embeddings for recognizing clinical text inference?

The BioNLP research has always suffered from the lack of labeled corpora due to privacy issues and increasing the costs of expert annotations. The existing datasets are usually modest in size and do not fit the requirements to serve as benchmarks for end-to-end NLP tasks (Lourenço et al., 2008). The ACL-MEDIQA challenge addresses the above limitations through three proposed tasks: Natural language inference, Recognizing question entailment, and entailment-based question answering (Ben Abacha, Shivade, & Demner-Fushman, 2019b).Whereas RQ5 shows which sentence representations methods are the most promising to model the medical knowledge of the clinical text, this chapter gives a more in-depth investigation of using the suggested methods in tackling the MEDIQA tasks. This research question mainly evaluates the generalizability of state-of-the-art text embeddings models to solve the text inference problem in the biomedical domain.

RQ7 — How can BioNLP models be employed to help domain experts in the information retrieval of evidence?

The primary goal of text mining in the medical domain is to assist in informing future decisions about the care of the patients through the analysis of textual data. While BioNLP has developed into an active research field, there still exist significant gaps between the BioNLP community and the medical community (Chapman, 2010). Among the challenges of clinical practice is the task of information retrieval that requires high levels of recall with acceptable precision. An example application is the conduction of systematic reviews, in which experts need to review a huge number of potential studies for inclusion, followed by abstraction and analysis of those studies (McGowan & Sampson, 2005). In the context of precision medicine, systematic reviews are a powerful tool that facilitate the decision making for clinicians by comparing the efficiency of medical tests or interventions with respect to the patient data (Boca, Panagiotou, Rao, McGarvey, & Madhavan, 2018). Therefore, this final research presents a case study that employs the previous research efforts in medical NLI to assist experts in judging the relevance of published literature for updating systematic reviews more efficiently.

1.4 Research Framework

For the past 30 years, Design Science Research (DSR) has been part of the engineering and information systems research (Gregor & Hevner, 2013). The DSR framework can be described as an outcome-based research paradigm that pursues creating novel information technology (IT) artefacts. The field of information systems (IS) is one of the many problem domains for design science research. Generally, design science improves our understanding by refining the search among probable constructs and components in order to develop the artefacts (Elragal & Haddara, 2019). It brings both practical as well as theoretical rigor to information systems research. According to Rai (Rai, 2017), there exist identifiable genres of design science research that vary according to the problem addressed, the types of artefacts, the search processes of creating and refining the artefacts, and the set of knowledge contribution. Our research falls in the *Computational Genre*. As implied by its name, output artefacts include representations, algorithms, analytics methods, human-computer interaction innovations, computer-assisted tools, and applications. This class follows an interdisciplinary approach to produce artefacts. In this thesis, the original Hevner three-cycle framework is adopted (Hevner, 2007; Hevner, March, Park, & Ram, 2004) as illustrated in figure 1.2. The framework is further specified to cater for its utilization of text mining for precision medicine. Additionally, the figure shows how the aforementioned research questions are mapped onto the framework, where the contribution of each research question in the context of the entire dissertation can be seen. To provide answers to RQ1–RQ7, we make use of several different research methods, both of a quantitative and a qualitative nature available within the framework. The rest of this section gives a summary of each of the three main cycles of Hevner's model, namely the cycles of *relevance, rigor*, and *design*. It also gives a brief description of the research methods used to accomplish this work. Although this list does not cover every research method that was employed in this dissertation, the most prominent ones are included. Table 1.1 shows the empirically addressed research questions, research methods, and data collection methods.

1.4.1 Relevance Cycle

With the desire to maximize the influence of their work, researchers set as a primary goal to meet the expectations of stakeholders faced with practical problems. The relevance cycle hence begins with identifying a problem or an opportunity from the field and, at the same time, set the criteria that will validate the research results. Once the research is complete, these results will be evaluated according to their ability to solve the problem. If these results are not satisfactory, either because they do not respond to all requirements, or because their use is limited, then a new iteration of this loop is recommended until a result is reached that satisfies the validity criteria initially issued.

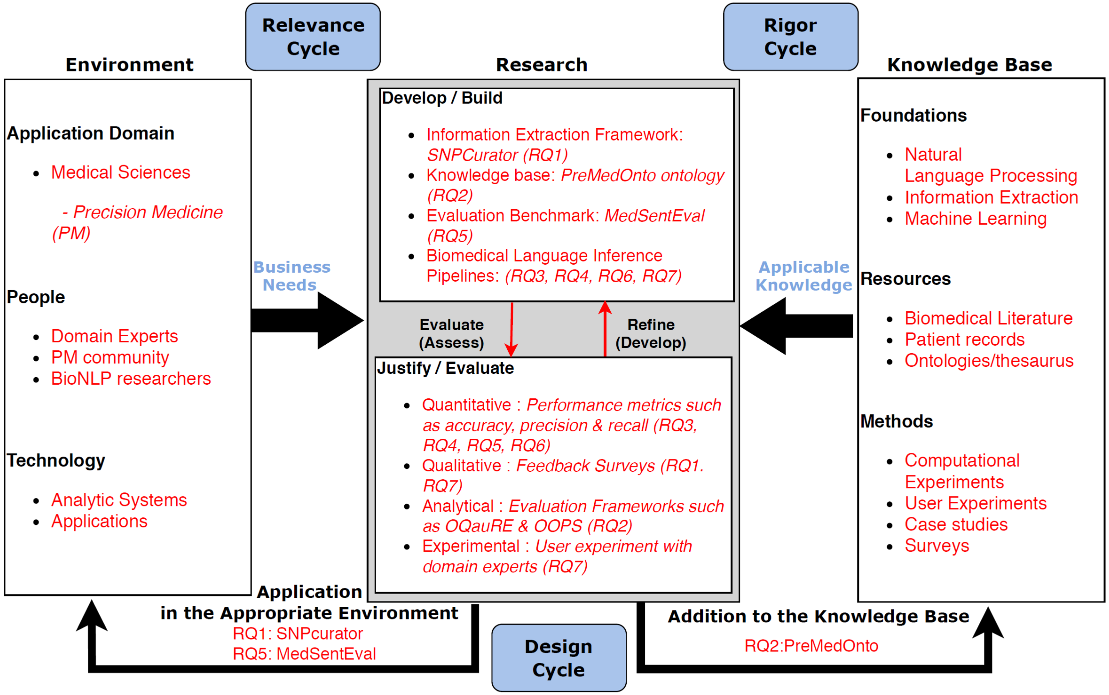

Relevance Cycle
Rigor Cycle
Environment
Research
Knowledge Base
Application Domain
Medical Sciences
- Precision Medicine (PM)
People
Domain Experts
PM community
BioNLP researchers
Technology
Analytic Systems
Applications
Business Needs
Develop / Build
Information Extraction Framework: SNPCurator (RQ1)
Knowledge base: PreMedOnto ontology (RQ2)
Evaluation Benchmark: MedSentEval (RQ5)
Biomedical Language Inference Pipelines: (RQ3, RQ4, RQ6, RQ7)
Evaluate (Assess)
Refine (Develop)
Justify / Evaluate
Quantitative : Performance metrics such as accuracy, precision & recall (RQ3, RQ4, RQ5, RQ6)
Qualitative : Feedback Surveys (RQ1. RQ7)
Analytical : Evaluation Frameworks such as OQauRE & OOPS (RQ2)
Experimental : User experiment with domain experts (RQ7)
Applicable Knowledge
Foundations
Natural Language Processing
Information Extraction
Machine Learning
Resources
Biomedical Literature
Patient records
Ontologies/thesaurus
Methods
Computational Experiments
User Experiments
Case studies
Surveys
Application in the Appropriate Environment
RQ1: SNPcurator
RQ5: MedSentEval
Design Cycle
Addition to the Knowledge Base
RQ2:PreMedOnto

Figure 1.2: Design research framework of this dissertation, along with a mapping of the seven research questions onto the corresponding components. Adapted from the information systems research framework by Hevner et al. (Hevner, 2007). Abbreviations: RQ = Research Question.

The relevance problem corresponds to the main research question of this dissertation as there lies great potential in the capabilities of text mining for advancing the field of PM. In this research, the environment refers to the medical health domain, in which domain experts (i.e. doctors, researchers, practitioners), healthcare organizations, and even patients act as stakeholders. In the first and last research questions, expert opinions regarding developed applications are taken into consideration through questionnaire-based evaluations. This serves in maintaining the feedback loop between phases.

1.4.2 Design Cycle

This cycle is of particular importance since it is at the heart of this DSR framework and controls how cycles interact. In this sense, the design cycle feeds initially on the first results of the loop relevance, i.e. the recognition of problems or opportunities to be seized. Building on March and Smith's work (March & Smith, 1995), Hevner et al. Hevner et al. (2004) introduces four types of artefacts: constructs, models, methods, or even instantiations. This cycle comprises two processes: design and evaluation, both complement each other to reach the final solution. Moreover, the design relies on research methods that take advantage of the domain knowledge (rigor cycle). In the evaluation cycle, artefacts must be assessed in terms of their usefulness (relevance cycle). In this perspective, the design loop is iterative.

1.4.3 Rigor Cycle

The rigor cycle of the research relies on the ability of the researcher to select and apply, from the knowledge base, theories, and methods how to design and evaluate the desired artefact. The knowledge base might be foundations, resources, or methodologies. This research lies in the foundations of, mainly, NLP and information extraction. In the context of the medical domain, the knowledge comes from structured sources such as ontologies, biological databases, and medical thesauri. Much of the clinical decision-making process is also guided by text-based processing of unstructured text such as published literature, clinical notes, and semi-structured data such as metadata. A measure of rigorousness is the ability to add to the knowledge base at the end of the research. The second research question contributes to the knowledge base with the publication of an ontology for precision medicine. The methodologies applied throughout the thesis were mostly quantitative, though qualitative methods were used as well. Below is a summary of the research methods explored: Experiment, Expert Survey and Case study.

Experiment

One of the founding quantitative methods in research is the experiment. Experiments are usually conducted in a controlled environment to validate or refute a hypothesis by measuring the influence of variables (factors) on the outcome (Seel, 2012). A computational experiment concerns itself with the theoretical analysis and empirical testing of a computational method, such as approximation or optimization algorithms. Such a purpose can include a demonstration of known truths, validating hypotheses, and examining the performance of something new. Typically, the effects and influences of controllable variables (factors) within an experiment are measured and studied on some phenomena. Within the computational context, experiments aim at theoretical analysis or testing the performance of a new computational artefact. In that setting, the experiment outcome is the performance of the proposed algorithm. This type is relevant to our research in text mining when an algorithm is applied to a dataset, resulting in empirical data that can be further analyzed or interpreted in a scientific context. Another type of experiment could be employed to analyze human behavior when confronted with information systems. This type is more traditional as it involves human subjects (Wohlin et al., 2000; Zelkowitz & Wallace, 1998). From a design science perspective, traditional experiments involving human subjects fit particularly well in the evaluation phases of the process. Computer experiments can be used in the design phases, while user experiments are employed to, obtain a more qualitative assessment of the developed artefact (Tedre & Moisseinen, 2014). Moreover, an experiment should follow the necessary methodological rigor to guarantee its validity. In this work, we discuss many standards of rigor for experimental research, such as external validity or generalization of models to other contexts, as illustrated in Chapter 6. We also report any experiment bias, when applicable, that causes limitations and skewness to our experimental results and reported findings. As shown in table 1.1, all chapters employ an experiment as one of the research methods. Chapters 2 to 7 use computational experiment while Chapter 8 uses a user experiment.

Expert Surveys

Surveys are the standard method to measure the assessment of people's thoughts, opinions, and feelings by means of an online or paper-based questionnaire or through a personal interview survey. Originally, they are more predominant within social sciences but have been widely adopted in the information science domain as well (Wohlin et al., 2000). In information science, surveys can be used for descriptive, explanatory, or explorative purposes and can be conducted before or after introducing a technique or tool (Ponto, 2015). Accordingly, within the design science paradigm, surveys are most suitable in the objectives-gathering and evaluation phases of the design science research process. As shown in table 1.1, chapters 2 and 8 employ surveys as one of the research methods.

Case Study

The case study research has evolved over the past few decades as a useful tool for investigating trends and specific situations in many disciplines. It is widely applicable in many domains, including social sciences, psychology, and information science. This methodology represents a research strategy to investigate theoretical models within a real-life context . By definition, a high degree of realism is achieved at the expense of the control levels (Runeson & Höst, 2009). Case studies align with the objectives of the design research framework as they aim at understanding a problem's specifics, determining a solution's objectives, and demonstrating and evaluating artefacts (G. Thomas, 2011). According to Yin (Yin, 2017), the case study research could be divided into holistic or embedded cases. In the first category, the case is studied as a whole, while in the second, multiple units of analysis are studied within a case. They are both suitable for inclusion in a DSR-based research according to the definition of the context and research goals. Finally, a case study may contain elements of other research methods, i.e. a survey may be conducted within a case study. As shown in table 1.1, chapters 2 and 8 employ a case study as one of the research methods.

1.5 Dissertation Outline

The research questions RQ1–RQ7 presented in Section 1.3 are investigated in Chapters 2–8 of this dissertation, with each research question corresponding to a single chapter. Each of these seven chapters are written as papers, published in proceedings of scientific conferences or in scientific journals.

Chapter 1 — Introduction describes the motivation and objectives of this research, along with the definition of Precision Medicine, the big data explosion in the medical research domain, and the computing research sub-domains relevant to this dissertation. We introduce the main overarching research question, seven specific research questions, and research methods that are used to an-

Table 1.1: A summary of the addressed research questions, research methods, and data

Chapter	**RQ**	**Research Method**	**Corpora**
2	RQ1	Computer Experiment Expert Survey Case Study	PubMed articles
3	RQ2	Computer Experiment	PubMed abstracts Ontologies (*NCIT,MeSH,IOBC*)
4	RQ3	Computer Experiment	PubMed abstracts (*ManConCorpus* dataset)
5	RQ4	Computer Experiment	PubMed abstracts (*ManConCorpus* dataset)
6	RQ5	Computer Experiment	PubMed abstracts Patient Records Patient Tweets
7	RQ6	Computer Experiment	Patient Records (*MedNLI* dataset)
8	RQ7	User Experiment Expert Survey Case Study	Cochrane Reviews

swer these questions. Finally, the dissertation outline which describes how the remainder of this dissertation is structured.

Chapter 2 — Literature Mining of Enriched Genetic Associations explores how NLP techniques could be combined to help experts and healthcare professionals in finding relevant information quickly and effectively. For this purpose, we introduce *SNPcurator*, an information extraction platform dedicated to researchers interested in GWA studies. It automatically extracts single-nucleotide polymorphism (SNP) associations of any given disease and returns a tabular list of relevant SNPs and the related statistical and demographic data. The online tool allows for query customisation, results sorting, and information expansion. The chapter gives a detailed description of the proposed framework and each available component. It also demonstrates the platform's effectiveness by comparing two case studies to existing gold-standard catalogs and by assessing its applicability through expert surveys.

Published as — Tawfik, N. S., Spruit, M. R. (2018b) The SNPcurator: Literature mining of enriched SNP-disease associations. *Database, 2018.*

Chapter 3 — Structuring the Precision Medicine domain To provide a coherent solution for structuring the PM knowledge, without duplication or conflict, we construct an ontology reuse framework. The framework relies on the assumption that reusing content would guarantee a consistent representation of the domain knowledge given the quality of the source ontology. Ontologies are also very often linked to the concept of metadata. Metadata are descriptive data that is associated with primary data in order to add information. In the PM context, this is evident as many general, investigations, diagnostics, and treatments' terminologies overlap considerably among different other domains such as oncology, for example. The related terms included in gold-standard ontologies can then be imported directly. A two-fold evaluation process was carried out to assess the newly developed Precision Medicine Ontology *PreMedOnto* for both the design correctness and quality aspects.

Published as — Tawfik, N. S., Spruit, M. R. (2019a). PreMedOnto: A Computer Assisted Ontology for Precision Medicine. In *International Conference on Applications of Natural Language to Information Systems* (pp. 329–336). Springer, Cham.

Chapter 4 — Highlighting Contradictions in Evidence-Based Medicine introduces a two-phase model to automatically extract clinical findings from biomedical abstracts and predicts existing conflicts among them with respect to a query. Information is extracted from abstracts as it gives a concise summary of the research conclusion without redundancy. For the first phase, the model combines semantic and domain-based features to enhance the claim detection process. The second phase relies on a Support Vector Machine (SVM) that integrates negation, antonyms, and alignment scoring to detect contradiction. The model is validated on *ManConCorpus*, a corpus related to the cardiovascular domain consisting of 24 different topics with a total of 259 abstracts.

Published as — Tawfik, N. S., Spruit, M. R. (2018a). Automated Contradiction Detection in Biomedical Literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 138–148). Springer, Cham.

Chapter 5 — Biomedical Textual Inference goes beyond chapter 4's objective in predicting contradictions to the more inclusive text entailment problem. The same dataset *ManConCorpus* was used after enriching its content to fit the standard NLI benchmark datasets. A comparison of several machine learning techniques is performed with the objective of achieving the best performance in recognizing inference between two input sentences. Hand-crafted features and dedicated sentence encoders were both used to represent the data. The techniques include hand-crafted features such as *string-based features, negation, polarity*, and sentence encoders such as *Universal Sentence Encoder (USE)* and *InferSent*. Subsequently, we employ several classification models, including methods such as the Random Forest and XGBoost algorithms and neural models such as Dense Neural Networks (DNN) with varying numbers of hidden layers. As expected, deep learning models outperform simple classification algorithms by a clear margin. Moreover, we investigate the benefits of incorporating traditional features with deep learning networks to boost performance. Overall, results are in favor of the hybrid model with a small performance gain over the DNN model. These findings suggest that traditional ML and deep learning models are complementary. Their combination in an end-to-end model can enhance the learning process and improve the predictions on evaluation and test sets.

Published as — Tawfik, N. S., Spruit, M. R. (2019b) Towards Recognition of Textual Entailment in the Biomedical Domain. In *International*

Conference on Applications of Natural Language to Information Systems (pp. 368–375). Springer, Cham.

Chapter 6 — Biomedical Text Representations This chapter is an exhaustive comparison of state-of-the-art methods in sentence representations when applied to the biomedical domain. We collect 10 medical datasets that vary in size including binary and multi-class sets. The datasets belong to 5 different classification problems and cover a variety of NLP tasks: textual entailment, sentence classification, sentiment analysis, question answering, and semantic text similarity. The data sources are also diverse among the datasets with biomedical literature, patient-authored texts, and clinical notes. The embedding representations included static embeddings (GloVe and FastText), context embeddings (BERT, ELMo and, Flair), and dedicated-sentence encoders (USE and InferSent). For each technique, we experiment with models pre-trained on general and medical data. While no single embedding technique or pre-trained model can be identified as the best, we generally note that context-dependent models outperform their peers and that pre-training on medical language models benefits the classification performance. Our experimental results unveil a number of important observations to guide researchers in choosing the most suitable model according to their specific task type. We furthermore provide the full workflow through *MedSentEval*, as a Python toolkit, with all codes and scripts to re-produce the results and possibly, extend the evaluation by including other datasets and/or embedding models.

Published as — Tawfik, N. S., & Spruit, M. R. (2020b) Evaluating Sentence Representations for Biomedical Text: Methods and Experimental Results. *Journal of Biomedical Informatics*, 103396.

.

Chapter 7 — Clinical Textual Inference describes our participation in the ACL MEDIQA challenge in tasks 1 and 2. Partially based on the results of the previous chapter, we perform a rigorous and deeper evaluation of the best pre-trained models when applied to these datasets separately. In the first task, we exploit BERT embeddings, trained on clinical notes, biomedical articles, and computer science literature, on the MedNLI dataset. Our best submission relied on a three-level ensemble with 30 BERT base models and achieved an accuracy of 0.852 on the test set. The investigation shows that consistent improvement is achieved by implementing successive ensembling, i.e. multi-level ensemble classifiers. For the second task, the exploratory embedding analysis showed that employing the Universal Sentence Encoder even with a simple classification model yields good results outperforming the RQE dataset baseline.

> **Published as** — Tawfik, N. S., Spruit, M. R. (2019) UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain. In *Proceedings of the 18th BioNLP workshop and Shared Task* (pp. 493–499).

Chapter 8 — Computer-assisted Relevance Assessment of Literature In an attempt to engage the medical experts in discovering the potentials of Text Mining and Machine Learning, this work introduces the case study of simulating the process of updating systematic reviews (SRs). Updating SRs typically requires 6 to 12 months of effort to formulate search queries, collect data, assess the relevance to the topic, and satisfy the inclusion criteria. We designed and conducted a controlled user experiment to investigate whether assessing document excerpts supported by Text Mining suggestions, specifically medical language inference labels, can speed-up of the relevance assessment process and achieve high recall. To collect data for the experiment, we extracted 6 systematic reviews from the Cochrane Database of Systematic Reviews and included a total of 143 PubMed abstracts. The experiment included 22 participants with each review of the 6 SRs evaluated at least 3 times. The interface was built through the online research platform Gorilla (https://gorilla.sc/) and comprised of two sequential tasks, *Abstract-view* and *Sentence-view*, with a demographic and feedback questionnaires in the beginning and end of the experiment respectively. Findings show that the quality of the judging is not only as good as when shown full abstracts but also in most of the cases better in terms of accuracy and recall. The participants rated the system as highly usable, and would likely continue using a similar one for their daily search and curation activities.

> **Published as** — Tawfik, N. S., & Spruit, M. R. (2020a) Computer-assisted Relevance Assessment: a Case-study of Updating Systematic Reviews. *Applied Sciences*, *10*(8), 2845.

.

Chapter 9 — Conclusion provides answers to the research questions, based on the investigations in the individual chapters. We furthermore elaborate on the best practices-based guidelines for improving the reusability and transparancy of our investigations, and describe the limitations of this research, directions for further research, and close with personal reflections.

2 Literature Mining of Enriched Genetic Associations

The uniqueness of each human genetic structure motivated the shift from the current practice of medicine to a more tailored one. This personalized medicine revolution would not be possible today without the genetics data collected from genome-wide association studies (GWASs) that investigate the relation between different phenotypic traits and single-nucleotide polymorphisms (SNPs). The huge increase in the literature publication space imposes a challenge on the conventional manual curation process which is becoming more and more expensive. This research aims at automatically extracting SNP associations of any given disease and its reported statistical significance (P-value) and odd ratio as well as cohort information such as size and ethnicity. Our evaluation illustrates that SNPcurator was able to replicate a large number of SNP-disease associations that were also reported in the NHGRI-EBI Catalog of published GWASs. SNPcurator was also tested by eight external genetics experts, who queried the system to examine diseases of their choice, and was found to be efficient and satisfactory. We conclude that the text-mining-based system has a great potential for helping researchers and scientists, especially in their preliminary genetics research. SNPcurator is publicly available at http://snpcurator.science.uu.nl/.

This work was originally published as:
Tawfik, N. S., Spruit, M. R. (2018b) The SNPcurator: Literature mining of enriched SNP-disease associations. *Database, 2018.*

2.1 Introduction

Ever since its completion in 2003, the Human Genome Project has accelerated and encouraged research on decoding the genome structure and functionality, powered by the huge advances in the genotyping technologies. The main goal of genomic studies is to identify and reveal the genetic variations associated with diseases and its prevalence across different populations. Such studies contribute to more tailored detection, prevention and treatment of diseases which lay the groundwork for the era of personalized medicine (Agyeman & Ofori-Asenso, 2015). In the hunt for correlation between genotype and phenotype, single-nucleotide polymorphisms (SNPs) are considered genetic signatures to the majority of polymorphisms responsible for trait susceptibility (Myles, Davison, Barrett, Stoneking, & Timpson, 2008).

By definition, a SNP is a single base-pair (A, T, C or G) variation that occurs at a specific site in the DNA sequence. It does not directly cause a disease but increases the genetic predisposition of individuals towards a certain disease and can affect their responses to drugs and medications (Carlson, 2008). Currently, information on SNPs is available in databases such as the genome-wide association study (GWAS) Catalog (Welter et al., 2014), Gwas Central (Beck, Hastings, Gollapudi, Free, & Brookes, 2014), GWASdb (M. J. Li et al., 2015), mirsSNP (Bruno et al., 2012), GRASP (Leslie, O\textquotesingleDonnell, & Johnson, 2014). These resources are constructed and curated manually; however, and the richness of information specifically related to the clinical impact of SNPs is contained in free text in the form of biomedical publications (P. E. Thomas, Klinger, Furlong, Hofmann-Apitius, & Friedrich, 2011). The process of updating databases requires substantial human resources, financial support not to mention time (Welter et al., 2014). This imposes a challenge as the number of published studies is steadily increasing and hence the manual curation is proving more and more inefficient.

Text-mining tools have been employed recently to overcome the mentioned limitations and accelerate the curation process. Mutation finder (MF) extracts mutations through regular expressions while tmVar (Wei, Harris, Kao, & Lu, 2013) also extracts mutations based on conditional random fields. Open Mutation Miner (OMM) (Naderi & Witte, 2012) uses MF to recognize single mutations and extends its regular expression set to detect mutation series. The extractor of mutations (EMUs) (Doughty et al., 2011) detects mutations in text and links them to genes, proteins and diseases. The SNP Extraction Tool for Human Variations (SETH) (P. Thomas, Rocktäschel, Hakenberg, Lichtblau, & Leser, 2016) implements an Extended Backus–Naur Form and regular expressions with more emphasis on short sequence variations and SNPs. Disgent database (Pinero et al., 2015) lists results compiled from expert curated databases and enhances the results by incorporating the BeFree text-mining system (Bravo, Piñero, Queralt-Rosinach, Rautschka, & Furlong, 2015). Polysearch (Y. Liu, Liang, & Wishart, 2015) associates genetic variants to diseases and drugs based on their co-occurrence frequency in abstracts. Most of the above tools achieved high performance levels on different corpora. However, there is still a gap between the research community and the biomedical text-mining community. Yepes and Verspour compare in (Yepes & Verspoor, 2014) the performance of EMU, OMM, MF, tmVar and SETH intrinsically on the Variome corpus, and extrinsically on the COSMIC and InSiGHT database. The study also discusses the technical aspects related to using the tools; some of them require an intermediate to advanced level of programming knowledge to use them. Furthermore, the evaluation of the practical utility of the tools is not properly investigated nor how can they be adapted to fulfill a researcher's tasks efficiently.

In this article, we present the SNPcurator, a system more oriented towards information extraction specifically in the genome wide and candidate genes studies. The proposed model is constructed out of different natural language processing (NLP) modules to aid scientists in their search for relevant disease-associated SNPs through an intuitive web interface. It incorporates both syntactic and semantic methods to extract relevant information from PubMed abstracts such as cohort size and ethnicity, SNP ids and the reported results. The motivation behind this research is to create a publically available, scalable and fully automated extraction tool.

2.2 Materials and Methods

SNPcurator has an online web interface at `http://snpcurator.science.uu.nl/` and allows researchers to easily query and search diseases and provides a useful resource for overview and summarization of associated SNPs found in literature. Sample code and the files used for evaluation are also available for download. Figure 2.1 illustrates the system's overview and workflow.

2.2.1 Data collection

The first step is to identify genetic research studies from the complete PubMed repository without any limitation on citations counts, publishing journals or a certain time period. The NCBI Eutilies (`https://eutils.ncbi.nlm.nih`

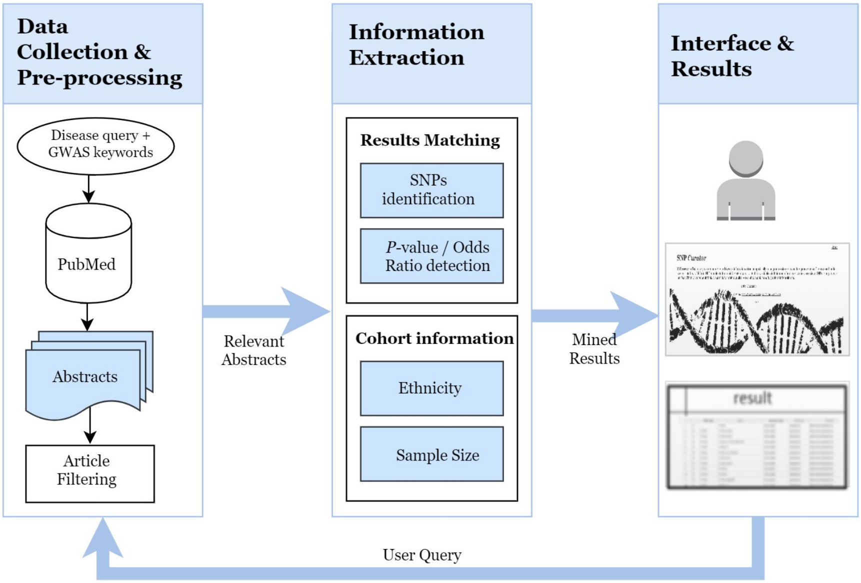


Figure 2.1: The SNPcurator workflow.

.gov/entrez/eutils/), in particular Esearch and Efetch, are used in conjunction with the BioPython module to query the PubMed database. All articles included in PubMed are associated with Medical Subject Headings (MeSH) terms (http://www.ncbi.nlm.nih.gov/mesh). MeSH is a controlled vocabulary used for indexing articles according to the unified Medical Language System meta-thesaurus which powers up the search capability of PubMed. SNPcurator combines the user input (the disease term in the current version) with relevant genetic search terms to construct the following PubMed query:

> *'{disease} AND("Polymorphism, Single Nucleotide"[Mesh Terms] OR "Genetic Predisposition to Disease"[Mesh Terms] OR "Genome-Wide Association Study"[Mesh Terms])'*

Esearch returns a list of IDs that matches the query search while *Efetch* returns data records for the retrieved ID list. Not all retrieved articles are included in the final search results; as the system only relies on the information found in abstracts and not the full-text article, all records with missing/incomplete abstract text or in a foreign language are excluded. The downloaded results will include literature with research on multiple genetic variants such as mutations or genes. To limit our search scope to SNP-association studies only, we apply a second filter to determine if the abstract text is relevant by checking SNP occurrences in text.

2.2.2 Information extraction

The filtered results are then pre-processed by several NLP modules. SNPcurator employs the publically available spaCy toolkit which is optimized for both accuracy and performance (`https://spacy.io/`). It accomplishes all necessary NLP tasks easily through its Python API. SpaCy is debatably one of the fastest publicly available parsers (Choi, Tetreault, & Stent, 2015). We make use of its sentence splitting, tagging, tokenization, name entity recognition and dependency parsing modules to extract SNP-disease pairs.

SNPs identification Scientists refer to SNPs in their research through unique ID numbers in a standard format assigned by the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/). The ID format differs between a newly submitted SNP (e.g.: ss28937569) or reference SNP cluster, refSNP (e.g.: rs28937569), where the latter is assigned to a submitted SNP after the sequence is aligned to its appropriate region (Kitts & Sherry, 2002). A regular expression formula _‘[rR][sS] []*[0-9][0-9]*’_ is used to identify the SNP identifier format and extract a list of SNP occurrences in the abstract text. This expression has previously achieved a 100% recall over a test set of 300 SNP mentions (Klinger et al., 2007). In our model, the latter formula has been improved to optionally accept characters G, C, T and A at the end of SNP mentions as authors tend to specify reference/alternative alleles or chromosome position in a non-standard format.

Results matching In the analysis of genetic variations, researchers follow exhaustive guidelines and conduct several statistical tests in order to report positive associations in GWAS which in return requires the definition of a known threshold. P-value, a parameter of statistical significance is used to determine the certainty of an association. It provides the probability that a given result from a test is due to chance with lower P-values giving higher probability of the association. Many values have been reported as a threshold to attain genome-wide significance (Panagiotou, Ioannidis, & others, 2012). Therefore, our model includes all studies even those with marginally significant and insignificant P-values. By displaying all results without filtering their significance, SNPcurator aims to help researchers rule out a given hypothesis or trigger more questions for further investigations.

To relate a reported value(s) to mentioned SNPs, the system first highlights one or more 'result sentence(s)' from the abstract. A 'result sentence' is identified as any phrase with P-value mention along with a numeric value. Each flagged sentence is first tokenized into words and then pre-processed to remove any unnecessary characters like quotes and brackets. There are also a number of naming patterns for reporting a P-value according to how it was calculated (e.g. *'P-combine', 'P-value', 'P-meta'*), which requires normalization to one single standard format that is easy to capture and extract. Following that, a set of predefined regular expressions are used to capture and recognize the format followed by floating, scientific or exponent number notations. However, since a sentence meaning relies on its semantics and structure, the coupling of P-value to its corresponding SNP is not the same for all sentences. The most common and straight forward way of stating results is by mentioning both SNP and P-value in the same sentence. In this type, the extraction of needed pairs ¡SNP, P-value¿ relies on the order as illustrated in the examples shown in Table 2.1. The system forms a pair by extracting the nearest P-value to the SNP mentions. Note that a detected P-value can only be coupled with a single SNP mention, and thus not considered when measuring the distance for the next SNP. This allows an accurate coupling when multiple values are involved such as in examples (2, 3 and 4). The same process also applies to the extraction of the odds ratio (OR) values reported in abstract text. In general, OR is a measure of the effect size in case-control study and in genomic-studies specifically, it denotes the probability of having the disease n individuals with and without certain genotypes of SNPs.

Patient information extraction The sample size included in the study is another key parameter when confirming the association with statistical confidence. For that purpose, we created two sets of keywords that are commonly used to describe both the patient cohort ('PatientKeywordSet') and the control group ('ControlKeywordSet'). The 'PatientKeywordSet' consists of words such as ['patient', 'case', 'subject'] and the 'ControlKeywordSet' includes words like ['control', 'normal', 'healthy']. The spaCy parser is then invoked to search for all numeric modifier (NUMMOD) dependencies found in the abstract text. The keyword sets are first compared against the head token of each candidate modifier and in the case of a no match; they are then compared against a list of two neighboring words of the candidate. Examples of control and patient group sizes extracted from evidence sentences are demonstrated in Table 2.2.

	PMID	Evidence sentence
1	21552555	We next examined obesity-related quantitative traits such as total body weight, waist circumference and waist to hip ratio, and detected genome-wide significant signals between waist to hip ratio and NRXN3 rs11624704, P = 2.67 ×10^{-9}, previously associated with body weight and fat distribution.
2	23143601	We identified three new susceptibility loci at 10q25.2 (rs7086803, P = 3.54 ×10(−18)), 6q22.2 (rs9387478, P = 4.14 ×10(−10)) and 6p21.32 (rs2395185, P = 9.51 ×10(−9)).
3	24880342	We identified large-effect GWASs for squamous lung cancer with the rare variants BRCA2 p.Lys3326X (rs11571833, OR = 2.47, P = 4.74 x 10(-20)) and CHEK2 p.Ile157Thr (rs17879961, OR = 0.38,P = 1.27 × 10(−13)).
4	21725308‘	The combined analyses identified six well-replicated SNPs with independent effects and significant lung cancer associations (P 5.0 x 10(-8)) located in TP63 (rs4488809 at 3q28, P = 7.2 10(-26)),$TERT - CLPTM1L(rs465498andrs2736100at$ 5p15.33, P = 1.2 x 10(-20) and P = 1.0 x 10(-27), respectively), MIPEP-TNFRSF19 (rs753955 at 13q12.12, P = 1.5 x 10(-12)) and MTMR3-HORMAD2-LIF (rs17728461 and rs36600 at 22q12.2, P = 1.1 x 10(-11) and P = 6.2 ×10(−13), respectively).

Table 2.1: Examples of (SNP, P/OR values) pairs extracted from evidence sentences

Because genetic variants often have markedly different frequencies across populations, the system is also able to extract the ethnicity and nationality of patients through the spaCy name entity recognition module. The *ents* attribute of the processed input abstract lists all entity objects found. An entity object is a combination of text, category and word position within text. We are particularly interested in the NORP entity which extracts nationalities, religious or political groups efficiently (where the latter two are irrelevant in this context). To limit the number of false positives, we also set the minimum word length for a nationality. If more than one nationality/ethnicity are found, the most frequent nationality mentioned would be reported as shown in Table 2.3

	PMID	Evidence sentence
5	2614121	We genotyped IL1B SNPs in a case-control study with 889 lung cancer cases and 1005 controls using the SNPscan Genotyping system.
6	25245582	A total of 169 ED patients (106 with anorexia nervosa (AN) and 63 with bulimia nervosa (BN)) and 312 healthy subjects were genotyped.

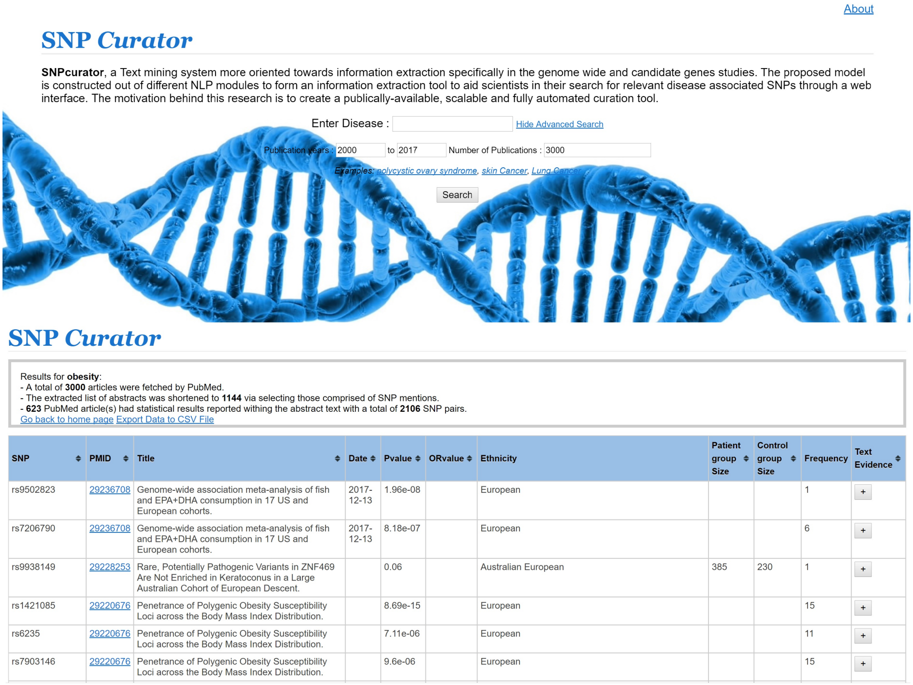
Table 2.2: Examples of control and patient group sizes extracted from evidence sentences

	PMID	Evidence sentence
7	22399527	We conducted a GWA study on MetS and its component traits in 4 Finnish cohorts consisting of 2637 MetS cases and 7927 controls
8	22399527	Therefore, we explored the association between the polymorphisms of CTSS and metabolic disorders in a Chinese Han population.

Table 2.3: Examples of nationalities and ethnicities extracted from evidence sentences

2.2.3 Interface

SNPcurator web service is implemented in Python and based on the Flask web development library (`http://flask.pocoo.org/`). Flask is a micro-framework for Python based on Werkzeug, the WSGI utility library and Jinja temple engine. It provides simple and flexible routing to build python-based web applications and also allows the integration of other python scripts. The disease query is passed from the front-end; the server initiates and displays the output from the text-mining module script. All the extracted information is displayed in a tabular format, where each row includes a single SNP item. The last column allows the user to check the result sentence from which the ¡SNP, P-value¿ tuple was extracted. The user is able to sort the results according to the P-value, OR value, group sizes, date of publication or simply alphabetical/numerical order of the PubMed or SNPs Ids. For example, by ranking results according to the ascending P-values, researchers can easily view the SNPs with the strongest association values that attained the genome-wide significance level. The web-tool is deployed on a server connected to the internet and can be accessed at `http://snpcurator.science.uu.nl/`, Figure 2.2 shows screenshots of SNPcurator's interface.



SNPcurator, a Text mining system more oriented towards information extraction specifically in the genome wide and candidate genes studies. The proposed model is constructed out of different NLP modules to form an information extraction tool to aid scientists in their search for relevant disease associated SNPs through a web interface. The motivation behind this research is to create a publically-available, scalable and fully automated curation tool.

Enter Disease : Hide Advanced Search

Publication years : 2000 to 2017 Number of Publications : 3000

Examples: polycystic ovary syndrome, skin Cancer, Lung Cancer

Results for **obesity**:
- A total of **3000** articles were fetched by PubMed.
- The extracted list of abstracts was shortened to **1144** via selecting those comprised of SNP mentions.
- **623** PubMed article(s) had statistical results reported withing the abstract text with a total of **2106** SNP pairs.

Go back to home page Export Data to CSV File

SNP	PMID	Title	Date	Pvalue	ORvalue	Ethnicity	Patient group Size	Control group Size	Frequency	Text Evidence
rs9502823	29236708	Genome-wide association meta-analysis of fish and EPA+DHA consumption in 17 US and European cohorts.	2017-12-13	1.96e-08		European			1	+
rs7206790	29236708	Genome-wide association meta-analysis of fish and EPA+DHA consumption in 17 US and European cohorts.	2017-12-13	8.18e-07		European			6	+
rs9938149	29228253	Rare, Potentially Pathogenic Variants in ZNF469 Are Not Enriched in Keratoconus in a Large Australian Cohort of European Descent.		0.06		Australian European	385	230	1	+
rs1421085	29220676	Penetrance of Polygenic Obesity Susceptibility Loci across the Body Mass Index Distribution.		8.69e-15		European			15	+
rs6235	29220676	Penetrance of Polygenic Obesity Susceptibility Loci across the Body Mass Index Distribution.		7.11e-06		European			11	+
rs7903146	29220676	Penetrance of Polygenic Obesity Susceptibility Loci across the Body Mass Index Distribution.		9.6e-06		European			15	+

Figure 2.2: The SNPcurator web interface.

2.3 Evaluation and discussion

2.3.1 Results

To assess SNPcurator's performance, we compared the information extracted from SNPcurator to data found in the GWAS Catalog for two queries: Obesity and Lung Cancer. According to cancer.org, Lung cancer is the second most common type of cancer that affects both men and women (combined) and it is the first cause of cancer death. Although obesity is not in itself a direct cause of death, overweight is a major risk factor of several diseases leading to death.

However, it is also considered one of the top preventable diseases. Therefore, by identifying individuals with increased risk of obesity, an early treatment plan would help to avoid mortal consequences (*Obesity and overweight Fact Sheet*, 2017).

SNPcurator was able to extract 1422 associations while the GWAS catalog shows a greater number of results with evidence of 1887 associations for the obesity trait. For the lung cancer disease, SNPcurator shows 620 associations versus 629 found in the GWAS catalog. This was expected given that GWAS catalog is manually curated from full-length articles while SNPcurator tool results are limited to the analysis of abstracts and titles only. Nevertheless, SNPcurator achieves a similar number of associations. It is important to note that SNPcurator results will almost certainly include false positives but as mentioned previously, reporting an association relies on a set of standards and rules that differ from one database to the other. SNPcurator results leave such evaluations to the user according to the information extracted. By observing the top 25 results when ordering associations in ascending order of their P-values, SNPcurator shows evidence for SNPs (rs8102683, rs465498 and rs12914385) and SNP (rs11127958) for the lung cancer and the obesity queries, respectively. Even though the GWAS catalog curation constraints might be the reason why these records are not listed in the catalog, we were still able to confirm these associations through other resources (M. J. Li et al., 2015).

To further investigate the text-mining module performance, we compared the results from SNPcurator to associations derived from abstracts and titles only in the GWAS catalog without considering data from full text. The catalog cites a total of 37 articles for the obesity disease, only a subset of 22 articles were selected. A chosen article must include both the SNP id and the corresponding P-value in its abstract-text. SNPcurator was able to replicate 21 associations while 9 associations were missed or not properly extracted. For lung cancer search; 25 associations were matched from a total of 34 associations reported in GWAS catalog extracted from 28 articles. On average, SNPcurator was able to replicate around 70% of the associations.

The performance of the information extraction component Recently, SNPPhenA, a new corpus of SNPs and Phenotypes associations extracted from GWA studies was published online (`http://nil.fdi.ucm.es/?q=node/639`). The corpus was constructed by querying the GoPubMed (`http://www.gopubmed.org/`) with 20 popular SNPs fetched from SNPedia. The original 20 SNPs names were used as seeds for the abstract collection process that resulted in 360 abstracts with 875 distinct SNPs. The novelty of the SNPPhenA corpus relies in ranking the associations by classifying them into three classes: positive, negative and neutral. The associations were manually annotated by two experts in the genetic fields and in case of any contradictory results; the verdict of a third annotator was taken into consideration. Moreover, a confidence level of positive associations was manually extracted based on the strength of the linguistic correlation between SNPs and disease mentions in the abstract, they were categorized into weak, moderate and strong. More details on the annotation process and the corpus statistics can be found at (Bokharaeian, Diaz, Taghizadeh, Chitsaz, & Chavoshinejad, 2017).

To our knowledge, SNPPhenA is one the first datasets to include the degree of certainty and confidence of associations instead of only reporting binary relations that simply include association or no association between SNPs and disease. Despite the relevance of the dataset to our work, the corpus authors relied on linguistic information, negation, modality markers and neutral candidates to label associations. In our approach, we determine the significance of associations in terms of biomedical statistical tests and study size. It is worth mentioning that P-values were regarded as an extra factor by the annotators of SNPPhenA when identifying the confidence levels of reported associations.

To properly evaluate our model, only a subset of the corpus SNPPhenA_mod was used. All records of the original corpus with no p-values/OR values reported in the abstract text were excluded and not considered when building the new corpus. The modified corpus, SNPPhenA_mod, consists of 120 abstracts with 166 key sentences and a total of 331 SNP-Phenotype association candidates with 160 distinct SNP identifiers. The new corpus was constructed following the manual annotation of associations by a biological expert, itis available for download in XML format from the about page. We evaluated the performance of extracting both (¡SNP, P-value¿) pairs in terms of precision, recall and F-measure. The model achieved 81%, 86% and 83%, respectively for each metric. What mostly affect the system sensitivity is the false negative (FN) cases that represents missed associations that were not detected by our system. Missed associations occurred due to failing to detect the P-value correctly or failing to link the correct P-value to the correct SNP. Failing to detect the correct P-value happens when values are reported in ranges instead of a single value. Another limitation of the system that it can only extract associations in the same sentence, i.e. both SNP and P-value are mentioned in the same sentence. It also assumes that one P-value applies to multiple SNPs if there is a single P-value mention with multiple SNPs mentioned in the sentence which in some cases results in False positives.On the other hand, failing to couple SNPs to their corresponding P-values contributes to a larger portion of FN and also to FP. Table 2.4 below illustrates some sample cases where SNPCurator failed to extract the correct associations.

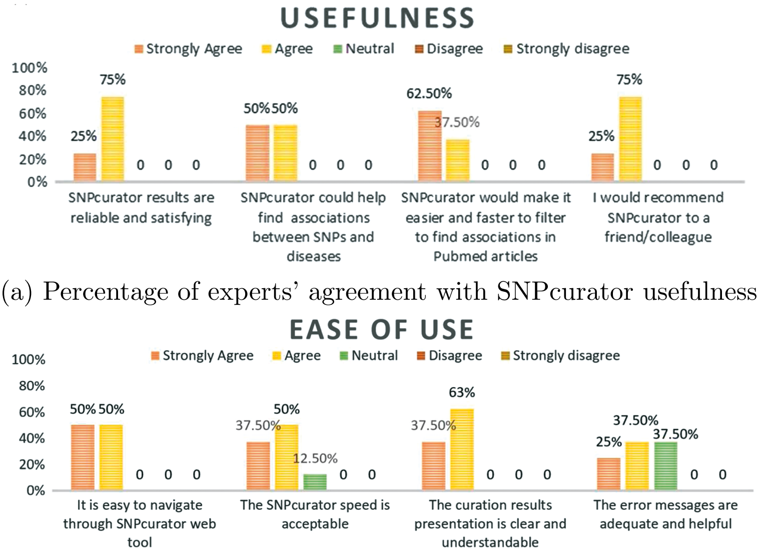
PMID	Evidence sentence	Reason
22914670	Two SNPs rs2656069 and rs10851906 in IREB2 were associated with COPD P = 0.045 and 0.032	Failure to detect the second P-value correctly led to missing the second associations
23065249	The objective of this study was to investigate the coding region polymorphisms S19W (rs3135506) and G185C (rs2075291) and the promoter region polymorphism 1131T >C (rs662799) of the APOA5 gene as risk factors for ischemic stroke in Turkish population. 19W allele frequency was 0.090 in stroke patients and 0.062 in controls P = 0.191.	The authors used different terms for identifying both SNPs in question (S19, G185).
18820697	rs5770917, a SNP located between CPT1B and CHKB, was associated with narcolepsy in Japanese (rs5770917[C], OR = 1.79,combined P = 4.4 10(7)) and other ancestry groups (OR = 1.40, P = 0.02).	The annotators matched SNP rs5770917C to the P-value while the system detected SNP rs5770917.
17383819	Significant association was detected at rs2254298 (P = 0.03) but not rs53576.	The system matched both SNP mentions to the P-value.

Table 2.4: Failed extraction cases

External evaluation Singhal et al. discuss in (Singhal, Leaman, et al., 2016) the need to advance biomedical sciences through text mining by empowering the roles of stakeholders involved (researchers, publishers and experts). Domain experts can evaluate the mined results and provide requirements, comments and guidelines for future improvements. For that reason, we presented the SNPcurator website to a number of interested scientists to solicit their feedback through an online survey. All eight participants have doctoral degrees in genetics or biosciences and are currently involved in mutations and polymorphisms research. Their years of experience in the genetic field varied; one participant was a full professor, four lecturers and two assistant lecturers and a genetic engineering expert from industry. Three participants were affiliated with the genetics department in the Suez-Canal Faculty of Medicine, while two were affiliated with the genetics department in Alexandria faculty of Medicine, two with the Medical research institute in Alexandria and one with Molecular Medicine and Tissue Culture sector of the European Egyptian Pharmaceutical Industries.

The survey followed the original Davis technology acceptance model guidelines (Davis, 1989) by addressing its two main aspects; the perceived usefulness of the system and its ease of use. Moreover, it collected suggestions for improvements as well as their opinions on the context in which they envision using the tool. Participants were encouraged to try different diseases of their choice and evaluate the extracted results. The overall system performance was satisfactory to all of the users; Figure 2.3 demonstrates users' responses to the questionnaire. Most of them agreed that SNPcurator would be useful for conducting a general preliminary research, studying SNPs variations among populations' comparisons or highlighting related studies for further readings.

Some participants pointed out that the initial results were limited and some known associations were missing despite their mention in multiple citations. For that we increased the number of papers downloaded from PubMed through eUtils to 3000 instead of only 1000. This resulted in longer loading time but we asked them to repeat the experts search query and this time, the missing associations were present. They also suggested to extract more data from the literature to enrich the records such as minor allele frequency and phenotypic effect. However, this information is usually found in the full-text and not abstract and hence could not be implemented for now. Another suggestion was to enable users to search by SNPs not just disease, this would require us to implement a SNP-Trait association extraction module, which we intend to pursue in future work.



(a) Percentage of experts' agreement with SNPcurator usefulness

(b) Percentage of experts' agreement with SNPcurator ease of use.

Figure 2.3: External evaluation results

2.4 Conclusion

In this article, we presented SNPcurator, a text-mining web-tool for automating the curation process of SNPs-trait information from GWASs. The system efficiently extracts associations and matches them with their statical significance parameters (P-value and OR) reported in abstracts; moreover it extracts population-related information such as the cohort sizes and ethnicity. To illustrate the tool's usability, we compared the results of two sample queries as opposed to the manually curated database. The system was able to report a comparable number of associations and also recreate a number of existing associations found in GWAS catalog. The system was also able to identify new associations not found in the catalog that might be interesting for experts to further investigate. Furthermore, to evaluate its usefulness and ease of use in daily research practice, a group of experts used the tool to search for any disease of interest; their opinions were very encouraging and confirmed the potential for SNPcurator. The tool was proven scalable and robust by applying it on the whole PubMed repository, and increasing the Faded articles from

1000 to 3000 per query. A main limitation was the analysis of abstract text only, we believe more accurate data would be extracted from full-text articles.

We acknowledge the fact that text-mining systems would never step up to the level of sophistication of researchers when reading and analyzing an article and hence would never fully replace human curation. However, it is also almost impossible to keep up with the fast publication rates of scientific research today. The speed and satisfying results of SNPcurator may be very beneficial to accelerate that process.

2.5 Future Work

Results have shown a difference between the amount of associations reported in abstracts and those reported in the article text itself. In the future, we aim to expand the text-mining scope to full-length articles and also any extra data provided by authors to further improve the system's performance. This would allow to extract more information regarding associations and also detailed cohort descriptions. Another potential add-on to the system would be a disease–SNP relation extraction module that allows an inverse query search by SNP id or even simply a set of PubMed articles ids. The system would also benefit from a general scoring scheme or a filter to rank the results not only based on the statistical findings but should incorporate as well the number of citations and the impact factors of the publications to indicate the credibility of the reported associations.

3 Structuring the Precision Medicine domain

This paper proposes an ontology learning framework that combines text mining, information extraction and retrieval. The proposed model takes advantage of existing structured knowledge by reusing terms and concepts from other ontologies. We further apply the methodology to create a detailed ontology for the emerging precision medicine (PM) domain by collecting a corpus of relevant articles and mapping its frequent terms to existing concepts. The resulting ontology consists of 543 annotated classes. The ontology was also tested for effectiveness by applying two evaluation frameworks to validate its design and quality. The results demonstrate that the ontology learning system is able to capture and represent the semantics of the PM domain with high precision and significance. Moreover, the computer-assisted construction process reduced dependency on expert knowledge.The developed *PreMedOnto* ontology could be further used to enhance the potentials of other NLP applications in the PM domain.

This work was originally published as:
Tawfik, N. S., Spruit, M. R. (2019a). PreMedOnto: A Computer Assisted Ontology for Precision Medicine. In *International Conference on Applications of Natural Language to Information Systems* (pp. 329–336). Springer, Cham.

3.1 Introduction

Ontologies are data models that transform domain's data into machine-readable representations to describe how a domain's information is organized. We adopt its original definition by Gruber as "An explicit specification of a conceptualization" (Gruber, 1993). By definition, they capture a wide variety of rich semantics by organizing knowledge into a hierarchy of concepts and relationships. It is considered one of the most reliable data representation models in today's semantic world, however, manual ontology development is an expensive task, both in terms of time and money. Ontology learning is the process of creating new ontologies from scratch whereas ontology population is concerned with augmenting existing ontologies with instances and properties. Both tasks require deploying efficient techniques to automatically process enormous amounts of domain-specific, unstructured resources. While the latter task is hard, the former task is particularly challenging as computer models must closely mimic domain experts in interpreting meanings for constructing the ontology (Buitelaar, Cimiano, & Magnini, 2005) and are usually accompanied by efficiency and precision issues.

An alternative to overcome such limitations is to take advantage of existing knowledge bases, as not only it would minimize the human factor, but it would potentially achieve better precision and reduce redundancy(Bontas, Mochol, & Tolksdorf, 2005). Reusing contents would also guarantee a consistent representation of domain knowledge given the quality of the source ontology. The practice is quite established as part of the Web Ontology Language (OWL) specification and is also supported by the Open Biological and Biomedical Ontology (OBO) Foundry (Ochs et al., 2017).

This study focuses on building an ontology for the precision medicine (PM) domain. The PM approach seeks to identify the best and the most effective practices for patients based on their genetic, environmental, and lifestyle factors. Although the concept has been around for many years, recently there has been an increase of public research funding and dedication to adopt the concept into practice versus the 'one-size-fits-all' method. Accordingly, there has been a substantial increase in the number of publications related to the PM concept (Yates et al., 2018). However, the PM domain lacks a clear and organized hierarchy of its general, investigations, diagnostics and treatments' terminologies. The main contribution of this research is the compilation and development of the precision medicine ontology (*PreMedOnto*). Such an ontology helps in defining and shaping the precision medicine domain and its related vocabulary which improves the understanding of the field.

3.2 Related Work

In the recent years, ontology has become a preferable way to represent biological data (Alobaidi, Malik, & Hussain, 2018). There is a great amount of published research in the ontology engineering field, however, our survey is only limited to ontology engineering models built for the medical domain. Casteleiro et al. was able to build an ontology for the sepsis disease from an unannotated biomedical corpus. Their model used Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), as well as the neural language models Continuous Bag-of-Words (CBOW) and Skip-grams (Arguello Casteleiro et al., 2018). They also exploited the same model to enrich the cardiovascular diseases ontology (CVDO) from PubMed articles. A reuse-based method was proposed by Gedzelman et al. to construct another ontology for cardiovascular diseases (Gedzelman, Simonet, Bernhard, Diallo, & Palmer, 2005) using UMLS and MeSH thesaurus. Cahyani and Wasito investigated the use of Ontology Design Patterns (ODP) to construct an Alzheimer's Disease ontology. Their model uses existing vocabulary and glossary to extract terms and relations from published articles and match them against the patterns (Cahyani & Wasito, 2017). Another Alzheimer's disease ontology was developed by Drame et al. (Dramé et al., 2014), they cluster bilingual terms from English and French corpora,according to the UMLS thesaurus, and align them by integrating new concepts.

In (Lossio-Ventura, Jonquet, Roche, & Teisseire, 2016), the authors propose a framework for updating existing medical ontologies. Their approach consists of 4 steps: extract relevant terms, apply machine learning techniques to infer polysemy, detect the concepts related to the term using clustering algorithms and finally, link terms to the exact positions in the ontology. Gao, Chen and Wang also suggested a model for extending ontologies (Gao, Chen, & Wang, 2018) and applied it to the PHARE ontology. Their research took advantage of PMC repository to train a word2Vec model and uses random indexing to enrich ontology labels. In (Kang, Fink, Doerfler, & Zhou, 2018), Kang et al. attempted to tailor the general adverse event ontology to build specific diseases ontology (DSOAE). They used design patterns and addressed the specifications needed for the chronic kidney disease by adding new classes, relations and properties. Another model was proposed in (Jiménez-Ruiz, Cuenca Grau, Sattler, Schneider, & Berlanga, 2008), where the authors reused the existing GALEN ontology to build a specific ontology for the juvenile rheumatoid arthritis disease. Their semi-automatic approach relies on extracting relevant parts of the old ontology and refine them to ensure consistency and safety so that the semantics of imported concepts are not changed. Amato et al. (Amato et al., 2015) populated an ontology constructed by a domain expert with RDF templates extracted from medical records. Sanchez and Moreno (Sánchez & Moreno, 2007) suggest a web based approach for building medical ontologies from scratch. It uses a set of user query words to collect web documents. Documents with the highest web search hit counts are considered valid taxonomic specialization for the domain. Named entities and verbs are then extracted to generate one-level taxonomy with general terms. The next stage is non-taxonomic learning where the extracted verbs are used as domain patterns and again used as web queries. Finally, the verb phrase is used to link each pair of concept. In (Alobaidi, Malik, & Sabra, 2018), Alobaidi et al. combined UMLS thesaurus and Linked Open Data (LOD) classes to identify medical concepts and associate them to their corresponding formal semantics. Shah et al. constructed a framework based on MetaMap and SemRep to reuse terms from SNOMED-CT ontology. They applied the framework to construct an ontology that combines the dental and medical domain to allow better reasoning over common knowledge (Shah, Rabhi, Ray, & Taylor, 2014).

3.3 Methods

3.3.1 Proposed model

Our ontology learning methodology is based on the concept of ontology reuse, where we adapt content from existing ontologies to model the PM domain. The model also relies on the assumption that the concepts that must be included in the ontology are mapped from the frequently mentioned terms present in the domain-specific data. And their co-occurrences frequency depicts the relations among them. To successfully achieve this goal, our proposed framework consists of 5 phases, figure 3.1 illustrates the overall learning process overview.

Knowledge Acquisition In our work, we used a publicly available list of PM keywords and synonyms constructed by conducting a systematic search through multiple web resources, including: academic, news and health websites. As this list is manually compiled and verified, we refer to it as the PM vocab. The list is divided into three categories: keywords and synonyms for personalized medicine, keywords and synonyms for personal genomics and keywords and synonyms for diagnostics, biomarkers and testing. More details on the creation of the vocabulary could be found in (Ali-Khan, Kowal, Luth, Gold, & Bubela, 2016). In this paper we only use the last category since

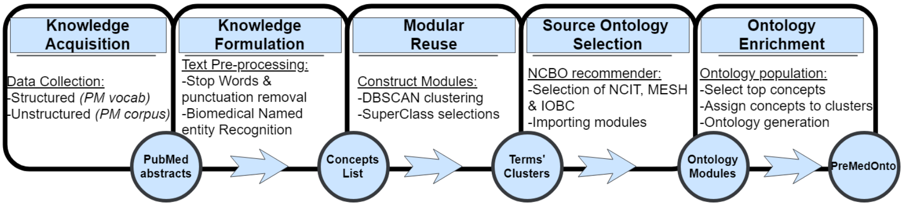


Figure 3.1: Overview of the ontology learning framework.

we aim at modelling the PM domain from a clinical and scientific point of view. In addition, we collected all titles and abstracts included in the PubMed repository discussing the PM concept. All articles included in PubMed are associated with Medical Subject Headings (MeSH) terms used for indexing articles. The search query used was ”precision medicine”[Majr], adding the [Majr] term next to the original query restricts the search engine to return citations where the PM concept is the major focus of the article. In scientific literature, medical terminology is usually used interchangeably to describe the same concept. The MeSH entry terms or cross-references ensure that closely related terms and synonyms are all included when querying a certain term. In our case, the entry list has other terms such as Personalized Medicine and Individualized Medicine. The collection process was conducted through the Bio Python package that connects to NCBI E-utilities to retrieve and download articles. The results are then filtered so that all records with missing or incomplete abstract texts or in a foreign language other than English are excluded. This resulted in a total of 5,206 articles that serve as the *PM corpus*.

Knowledge Formulation We preprocess all abstracts in the *PM corpus* to filter out stop words, symbols and punctuation. Due to the ambiguity of reporting biological or clinical results, MetaMap[1] was used for medical entity recognition. The output at this stage is a set of 6,832 distinct terms and concepts from the corpus. To guarantee precision, we do not map all terms extracted as this could lead to ambiguity and inconsistency in representing the domain knowledge. All terms mentioned more than once are ranked in descending order of their occurrence frequencies. Extracted terms are included only if their mention count exceeded a threshold. The threshold value is calculated as the weighted average occurrences of terms in documents to ensure that less significant words are removed.

[1] `https://metamap.nlm.nih.gov/`

Modular reuse In this stage, the *PM vocab* is used to create seed ontology modules where terms are mapped to a set of disjoint clusters. We started by analyzing the terms included in the *PM vocab* according to their relevance and commonness. We built a symmetric matrix of cosine similarity scores for every pair of word vectors that exist in the vocabulary. The word embeddings model was pretrained over a set of over 10 million biomedical articles from PubMed. The matrix was fed to a density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm implemented through the Scikit library. We opted for the DBSCAN clustering algorithm since it allows unsupervised learning over data and does not require the number of clusters a priori. This process created a total of 5 clusters. Following the creation of clusters, we rank all terms included according to their centrality and create one module per cluster. The top ranked concepts per cluster serve as the ontology super-classes. The original *PM vocab* set contained 100 terms that refer to diagnostic and testing procedures. Out of the 100 terms, only 73 were correctly clustered while 27 terms were regarded as noise by the clustering algorithm. Among the top candidate terms for each cluster, 25 were mapped as parent and child classes. Finally, we add all the non-used terms from the PM vocab to the list of concepts extracted from the *PM corpus.*

Source Ontology Selection It is critical to determine the correct ontology that can serve as the base of the newly developed *PreMedOnto*. The criteria of choosing the ontology include coverage, acceptance and semantic language used. The NCBO ontology recommender is employed to suggest the best ontology for each module over all 895 existing ontologies. To maximize the coverage factor, we opted for the ontology set option which returns the best set of combined ontologies. The weights configuration for the recommender scoring function was set to the default settings. The final ranking of ontologies to be reused was: National Cancer Institute Thesaurus (NCIT)[2] , Medical Subject Headings (MeSH)[3] and Interlinking Ontology for Biological Concepts (IOBC)[4]. From the selected ontologies, we import all candidate classes with their ancestors, and verify that all remaining concepts per cluster are included in the module as child nodes. All redundant concepts in the *PM vocab* are removed by checking synonyms of each imported class.

Ontology Enrichment In the final stage, each module is enriched by assigning relevant concepts extracted from the *PM corpus* in the knowledge

[2] `https://bioportal.bioontology.org/ontologies/NCIT`
[3] `https://bioportal.bioontology.org/ontologies/MESH`
[4] `https://bioportal.bioontology.org/ontologies/IOBC`

formulation phase. We first extract the Uniform Resource Identifier (URI) corresponding to each concept. The ontofox (Xiang, Courtot, Brinkman, Ruttenberg, & He, 2010) tool supports efficient ontology reuse by extending the Minimum Information to Reference an External Ontology Term MIREOT concept. The MIREOT approach favors selective class imports instead of importing the ontology as a whole. The ontofox web tool takes as input the base ontology, source terms URIs and parent classes URIs. It also allows users to choose the appropriate settings of the import process such as importing or omitting intermediate classes between input child and parent or deciding which annotation properties to return.

3.3.2 Evaluation

Assessing the ontology output is a key factor in all ontology learning techniques. Not only to ensure the ontology quality before referencing and adopting it in other semantics-aware applications, but also to highlight errors and shortcomings. There are two different evaluative perspectives: ontology quality and ontology correctness. In this research, we carried out a two-fold evaluation process to measure the effectiveness of the constructed ontology: the first experiment assesses the ontology design whereas the second computes multiple quality features. To detect any design error in *PreMedOnto*, we use OntOlogy Pitfall Scanner (OOPS) online tool (Poveda-Villalón, Carmen Suárez-Figueroa, Ángel García-Delgado, & Gómez-Pérez, 2009). OOPS evaluates an OWL ontology against a catalogue of common mistakes in ontology The tool produces a summary of all pitfalls found within the ontology with extended information on each and a label indicating its importance level. We also apply the ontology quality evaluation framework (OQauRE) (Duque-ramos, Duque-ramos, Fernández-breis, Stevens, & Aussenac-gilles, 2011) to validate the quality of classes and axioms in *PreMedOnto*. OQauRe is a quantitative method based on the original software product quality requirements and evaluation concept. The framework computes multiple quality characteristics including structure, quality in use, reliability, compatibility, maintainability, operability, functional adequacy, transferability, performance efficiency. The generated metrics are mapped to quantitative values ranging from 1 to 5 with 3 is the minimum score and considered as accepted.

3.4 Results

The final output of the ontology learning process is the *PreMedOnto* in the standard OWL format. A total of 543 classes imported from 3 medical on-

Table 3.1: Summary of the *PreMedOnto* metrics generated by the Protégé framework.

Metric		Metric	
Classes	543	Classes with a single child	111
Average number of children	3	Maximum number of children	90
Properties	10	Maximum depth	7

tologies. Table 3.1 provides a brief summary of some of its metrics.The ontology can be accessed, viewed and downloaded from `http://bioportal.bioontology.org/ontologies/PREMEDONTO`. The obtained results of evaluating *PreMedOnto* against the 41 pitfalls included in OOPS's catalogue, show that the ontology is free from critical and important pitfalls while there exist 3 cases of minor pitfalls. The former finding ensures the consistency and sustainability of the ontology, while the later might suggests corrections for better organization. The pitfalls detected are related to missing annotations, lack of connectivity and inverse relationship declaration. However, we find them irrelevant, as they do not threaten the functionality of the ontology. The second experiment provides quantitative indicators of the quality of *PreMedOnto*. The computed scores for structure, compatibility and maintainability metrics were 3.5, 4.2 and 4.5 respectively. The ontology has successfully passed the minimal level required and is considered above average in most characteristics. It is worthy to mention that each quality measure is also associated to multiple sub-characteristics and hence indicates multiple quality aspects.

3.5 Conclusions

PreMedOnto is an application ontology built for the precision medicine domain on top of gold standard biomedical ontologies. The ontology learning process involves mining the PubMed repository to extract domain specific abstracts and vocabulary as sources of data. The information gathered is clustered and outlined to determine main modules. It reuses terms and concepts from NCIT, MeSH and IOBC to construct the ontology hierarchy. The evaluations demonstrate that the ontology content is reliable and consistent. We also plan to add a possible extra experiment to validate the ontology utility and applicability in the PM domain. The intended experiment involves human validation of the ontology by medical experts through a survey of questions.

4 Highlighting Contradictions in Evidence-Based Medicine

Medical literature suffers from inconsistencies between reported findings that answer the same research question. This paper introduces an automated two-phase contradiction detection model that integrates semantic properties as input features to a Learning-to-Rank framework, to accurately identify key findings of a research article. It also relies on negation, antonyms and similarity measures to detect contradictions between findings. The proposed technique is implemented and tested on a publicly available contradiction corpus 259 manually annotated abstracts. The performance is compared based on recall, precision and F-measure. Experimental evaluations prove the utility of the model and its contribution to the contradiction classification and extraction task.

This work was originally published as:
Tawfik, N. S., Spruit, M. R. (2018a). Automated Contradiction Detection in Biomedical Literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 138–148). Springer, Cham.

4.1 Introduction

In the last decade, there was a substantial increase in the total number of medical research publications worldwide. Most of the literature publish results on the effectiveness of clinical interventions, and despite the similarity of the scientific experiment designs, not all outcomes are in agreement (J. P. Ioannidis, 2005). Whether published findings are consensual, complimentary, or contradictory facts, many of them get approved, updated or replaced accordingly (Prasad, Cifu, & Ioannidis, 2012a). Given the varying nature of published findings, it is difficult to fairly assess evidence-based knowledge within articles. More importantly, differences between research outcomes should be highlighted so that further studies do not build assumptions and/or conclusions on prior research that have since been disapproved, and are not valid anymore. In extreme cases, some published evidence-based facts even get reversed. Prasad et al. (Prasad et al., 2013), reviewed 363 articles published in one high impact factor journal investigating various established medical practices. While 138 (38%) confirmed the practices, 146 (40.2%) found them ineffective and 79 (27.3%) were inconclusive. For example, four studies contradicted the administration of the Aprotinin drug, widely used for treatment in post cardiac surgeries.

Such misinformation, or disinformation, create controversy that is important to researchers and practitioners interested in finding evidence-based answers to clinical queries; whether it is for the benefit of their patients or for the sake of conducting systematic reviews. It is also of great significance to both Comparative Effectiveness Research (CER) and the Precision Medicine (PM) communities. The comparative effectiveness research is interested in the analysis of medical interventions by comparing their benefits and drawbacks, to reach informed evidence-based decisions for a better clinical practice (Sox & Greenfield, 2009). While Precision Medicine also aims at improving the health care system, PM is different than CRM as it takes into account the genetic, environmental and lifestyle differences between individuals (Jameson & Longo, 2015). Highlighting different outcomes to the same medical practice supports the PM claims that there is no ”one-size-fits-all” treatment strategy.

However, with the high rate of growth in scientific publications, the task of finding answers, interpreting outcomes and validating them becomes tedious, exhausting and time consuming, even in a specific sub-domain. In result, several text mining tools and frameworks were built and employed to solve the information extraction problem, automatically or semi-automatically, for a variety of research applications. Biomedical text mining faces a number of challenges; the enormous number of existing publications, the unstructured nature of text, and most challengingly, the ambiguity of reporting biomedical or clinical results. Findings can be expressed in long, context-dependent sentences with the usage of a wide variety of terminology.

Contradiction Detection

in text is still a relatively new area of research. As in other Natural Language Processing (NLP) sub-domains, it requires a multidisciplinary approach involving text mining, sentiment analysis, opinion mining, knowledge retrieval and information extraction. This paper focuses on the problem of extracting contradicting findings in biomedical texts. In this context, we propose an automated contradiction detection framework that adapts and extends existing NLP tools. The proposed model takes advantage of a recently published corpus, constructed for the same purpose, to validate its accuracy.

4.2 Related Work

Despite the fact that more research has been conducted on text entailment rather than contradiction detection, the development of two contradiction corporas encouraged more research into the domain (De Marneffe, Rafferty, & Manning, 2008; Padó et al., 2008; Ritter, Downey, Soderland, & Etzioni, 2008). The corporas were based on direct negation and paraphrasing of sentences from the PASCAL Recognizing Textual Entailment (RTE) dataset (Harabagiu, Hickl, & Lacatusu, 2006). However, contradiction analysis remains a challenging task, mainly due to the different ways in which contradictions can appear (numeric mismatching, negations, contrastive sentences, etc.).

The importance of extracting contradictions has been exploited in other domains, most commonly in news and rumors text processing. Due to the generic type of negation found in normal text, it is difficult to adapt any of the former models to the biomedical domain directly. The language used to express biomedical facts is usually rich in clinical semantics and conceptual overlaps, and involves complex sentence structures. To the best of our knowledge, there has been minimal research conducted on the biomedical contradiction analysis.

In 2011, Sarafraz et al. (Sarafraz, 2012a) investigated both rule based and machine learning methods to identify negated molecular events through lexical, syntactic and semantic features, the model was evaluated on the BioNLP09 challenge corpus. Alamri et al. (Alamri, 2016a; Alamri & Stevensony, 2015) explored the use of four features: negation, directionality sentiment and uni+bi-grams combined with an SVM classifier, to extract contrasted findings reported in cardiovascular research literature. More recently, Preum et al. (Preum, Mondol, Ma, Wang, & Stankovic, 2017b) presented *Preclude*, a rule-based system that highlights conflicts in wellness advice, found in on-line health forums. The system constructs a polarity lexicon from verbs ,in the training set, and their synonyms using WordNet, for labeling actions found in text as positive or negative. In another attempt to discover the ambiguities in the biomedical literature, Silva et al. (de Silva, Dou, & Huang, 2017a) proposed an ontology-based system to extract inconsistencies found in miRNA research articles in the PubMed repository. The system relies on OLLIE "Open Language Learning for Information Extraction" framework to extract all relevant triples (subject, object, and relationship) from abstracts. Triple entries are then compared against each other to find inconsistencies, based on an *oppositeness metric* suggested by the authors.

4.3 Dataset

The lack of annotated data has led to the unavailability of comparison and evaluation of contradiction detection systems in the biomedical literature. However, this may change with the recent availability of Manual Contradiction Corpus (*ManConCorpus*), a corpora of contradictory research claims [1]. The corpus is constructed out of 24 systematic reviews on four important cardiovascular disease topics: Cardiomyopathy, Coronary artery, Hypertensive and Heart failure. Each review article is mapped to a closed PICO (Population, Intervention, Comparison and Outcome) question that could be answered only by Yes or No. The mapping process was conducted manually by a medical expert, after reviewing all research abstracts of studies included in the systematic review. These abstracts include research claims with answers to the questions. *A research claim* is a one-sentence summary of the research findings that the authors find important, either to affirm old information or to introduce new ones. Two annotators were asked separately to find one correct claim per abstract and label it *YES*, if it positively answers the question and *NO* otherwise. It is worth mentioning that despite the fact that multiple sentences in the abstract might hold the answer to the query, only the most informative one is chosen as per the annotator's opinion. The corpus has a total of 259 abstracts, out of which 180 introduce positive claims and 79 intro-

[1] Corpus available at `http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/resources/bio_contradictions/`

duce negative claims. All claims included in the corpus are either evaluative or causal. The former is an assessment of the biomedical concept presented in the research topped by a judgment, while the latter is a statement that describes the relation type between two concepts and whether one affects the other or not . More details on the annotation process and the corpus statistics can be found in (Alamri & Stevenson, 2016a).

In literature, there is no standard definition of 'contradiction', and it is usually task-dependent according to the nature of the contradiction instances. Therefore, we adopt the authors' definition of contradiction that better matches the corpus and human intuitions: "*Two texts, T_1 and T_2, are said to contradict when, for a given fact F, information inferred about F from T_1 is unlikely to be true at the same time as information about F inferred from T_2*". As per the definition, if both a positive and a negative claim answer the same query, they are considered contradictory as shown in the example in Figure 7.1.

4.4 Methods

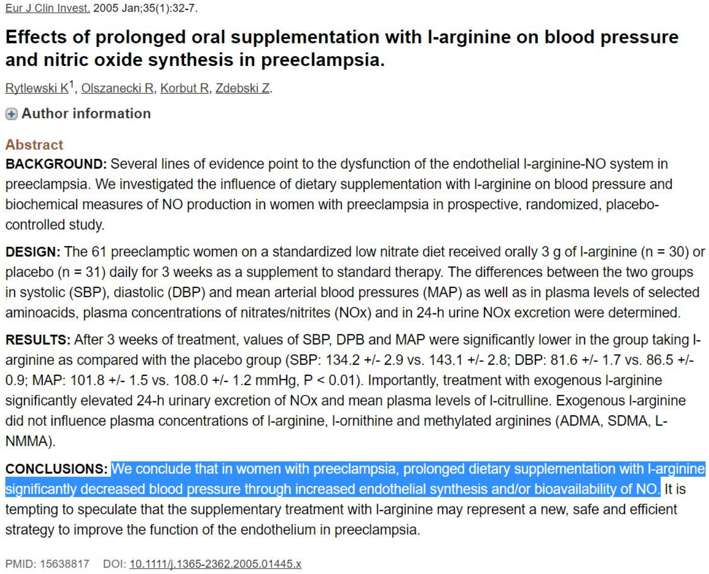
Identifying inconsistencies in text is a two-phase problem, claim retrieval and claim assertion. During the first phase, we need to identify potential sentences relevant to the query. In the claim assertion phase, we have to evaluate whether sentences infer text entailment or contradiction.

4.4.1 Identification of Abstract Claims

Finding relevant sentences that answer the query is a key component in the biomedical contradiction detection system, as the performance of the system is dependent on the accuracy of the extracted key phrase. Several methods and techniques have been employed for passage retrieval in general, and answer identification in specific. Nevertheless, it still remains a challenge in the biomedical language processing field, mainly due to the complex nature of the text. In this research, we address the claim extraction process as a ranking problem, where each sentence in the input text is scored according to its relevance to the query.

Input Preprocessing

We split all abstract text included in the corpus into sentences using The Natural Language Toolkit (NLTK). All sentences with less than three words are



Eur J Clin Invest. 2005 Jan;35(1):32-7.

Effects of prolonged oral supplementation with l-arginine on blood pressure and nitric oxide synthesis in preeclampsia.

Rytlewski K[1], Olszanecki R, Korbut R, Zdebski Z.

⊕ **Author information**

Abstract

BACKGROUND: Several lines of evidence point to the dysfunction of the endothelial l-arginine-NO system in preeclampsia. We investigated the influence of dietary supplementation with l-arginine on blood pressure and biochemical measures of NO production in women with preeclampsia in prospective, randomized, placebo-controlled study.

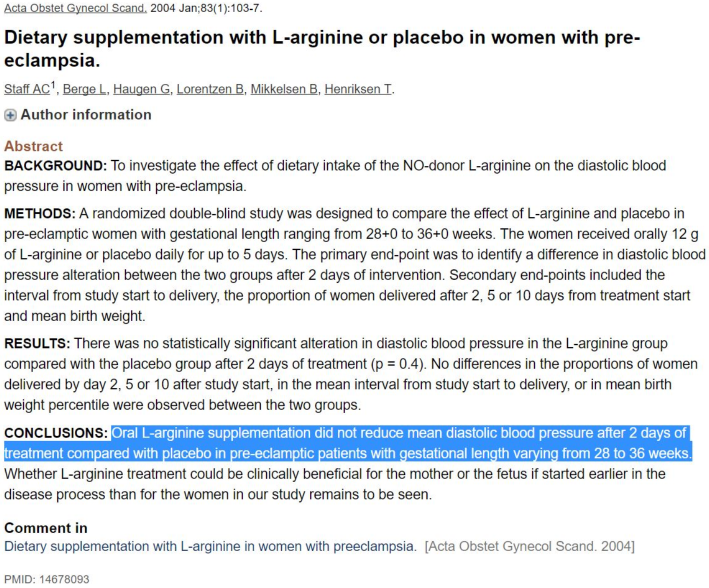
DESIGN: The 61 preeclamptic women on a standardized low nitrate diet received orally 3 g of l-arginine (n = 30) or placebo (n = 31) daily for 3 weeks as a supplement to standard therapy. The differences between the two groups in systolic (SBP), diastolic (DBP) and mean arterial blood pressures (MAP) as well as in plasma levels of selected aminoacids, plasma concentrations of nitrates/nitrites (NOx) and in 24-h urine NOx excretion were determined.

RESULTS: After 3 weeks of treatment, values of SBP, DPB and MAP were significantly lower in the group taking l-arginine as compared with the placebo group (SBP: 134.2 +/- 2.9 vs. 143.1 +/- 2.8; DBP: 81.6 +/- 1.7 vs. 86.5 +/- 0.9; MAP: 101.8 +/- 1.5 vs. 108.0 +/- 1.2 mmHg, P < 0.01). Importantly, treatment with exogenous l-arginine significantly elevated 24-h urinary excretion of NOx and mean plasma levels of l-citrulline. Exogenous l-arginine did not influence plasma concentrations of l-arginine, l-ornithine and methylated arginines (ADMA, SDMA, L-NMMA).

CONCLUSIONS: We conclude that in women with preeclampsia, prolonged dietary supplementation with l-arginine significantly decreased blood pressure through increased endothelial synthesis and/or bioavailability of NO. It is tempting to speculate that the supplementary treatment with l-arginine may represent a new, safe and efficient strategy to improve the function of the endothelium in preeclampsia.

PMID: 15638817 DOI: 10.1111/j.1365-2362.2005.01445.x

(a) PMID: 15638817 with assertion value YES



Acta Obstet Gynecol Scand. 2004 Jan;83(1):103-7.

Dietary supplementation with L-arginine or placebo in women with pre-eclampsia.

Staff AC[1], Berge L, Haugen G, Lorentzen B, Mikkelsen B, Henriksen T.

⊕ **Author information**

Abstract

BACKGROUND: To investigate the effect of dietary intake of the NO-donor L-arginine on the diastolic blood pressure in women with pre-eclampsia.

METHODS: A randomized double-blind study was designed to compare the effect of L-arginine and placebo in pre-eclamptic women with gestational length ranging from 28+0 to 36+0 weeks. The women received orally 12 g of L-arginine or placebo daily for up to 5 days. The primary end-point was to identify a difference in diastolic blood pressure alteration between the two groups after 2 days of intervention. Secondary end-points included the interval from study start to delivery, the proportion of women delivered after 2, 5 or 10 days from treatment start and mean birth weight.

RESULTS: There was no statistically significant alteration in diastolic blood pressure in the L-arginine group compared with the placebo group after 2 days of treatment (p = 0.4). No differences in the proportions of women delivered by day 2, 5 or 10 after study start, in the mean interval from study start to delivery, or in mean birth weight percentile were observed between the two groups.

CONCLUSIONS: Oral L-arginine supplementation did not reduce mean diastolic blood pressure after 2 days of treatment compared with placebo in pre-eclamptic patients with gestational length varying from 28 to 36 weeks. Whether L-arginine treatment could be clinically beneficial for the mother or the fetus if started earlier in the disease process than for the women in our study remains to be seen.

Comment in

Dietary supplementation with L-arginine in women with preeclampsia. [Acta Obstet Gynecol Scand. 2004]

PMID: 14678093

(b) PMID: 14678093 with assertion value NO]

Figure 4.1: An example of two contradictory claims found in literature that answer the query: *In women with pre-eclampsia, does treatment with L-Arginine, compared to a placebo, reduce blood pressure?*

considered an error of the splitting process, and thus eliminated. Afterwards, a set of potential claims that answer the query correctly is compiled for each abstract. As in any text mining application, the input text might be totally unstructured or semi-structured, and the same applies for literature abstracts. For slightly structured abstracts, i.e abstracts where text is divided into subsections such as Title, Introduction/Background, Methods/Aims, Results, and Conclusion, we take advantage of this information and include all sentences within the headings, Results and Conclusion, as candidate sentences. If the text is unstructured, all sentences in the second half of the abstract are included the candidate set, following the assumption that important findings are most probably reported by the end of the abstract. The candidate set is filtered out from any stop words, symbols and punctuations. All 24 PICO questions went through the same filtering process as the candidate sentence collection.

Feature Extraction

A fixed length feature vector representing sentences included in the candidate set is derived. These features combine both semantic and syntactic properties of the sentence. They capture relevance to the query, as well as relatedness to domain-specific concepts . In our model, we rely on easy-to-compute features, which have proven successful in other retrieval tasks. All of the following features are extracted for each of the candidate sentences.

1. *Sentence Length* The count of terms per sentence after removal of stop words.

2. *Sentence Location* The relative position of the sentence within the abstract as it highlights the importance of a sentence. The feature is calculated as the location divided by the total number of sentences. However, instead of using the original location of the sentence, we use its position in the candidate set.

3. *Term Overlap* This measures the number of terms that are found in both the query and the sentence, after removing of stopping words, and also stemming all terms using Porter stemmer (Porter, 1980).

4. *Synonyms Overlap* The fraction of overlap between the query terms and sentence terms or their synonyms fetched from WordNet.

5. *BM25 score* The Okapi BM25 framework is a Bag-of-Words model with a collection of scoring functions combined. For a query Q containing

n terms {q_1,q_2,q_3,....,q_n}and a sentence length S, the similarity score between$Q$ and S is calculated as

$$\text{BM25score}(S,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,S) \cdot (k_1+1)}{f(q_i,S) + k_1 \cdot (1 - b + b \cdot \frac{|S|}{\text{avgSl}})}$$

where IDF is the *inverse document frequency*, *avgSl* is the average sentence length and *k-b* are two tuning parameters set to 0.75 and 1.5 respectively in our implementation.

6. *Word Embeddings* Cosine similarity between query and candidate sentences' word vectors pre-trained using a set of over 10 million PubMed abstracts.

The first four features are a subset of the features suggested and used by Metzler and Kanungo's work on text summarization (Metzler & Kanungo, 2008). The last two features give more insight into the context of the sentences. In general, word embeddings are a vector representation of words co-occurrences that regard words as contexts, and hence gives a better apprehension of the word meanings. The vectors are generated from a large collection of data through Neural Networks. We take advantage of the word vectors, provided by the BioASQ challenge team (Ioannis Pavlopoulos, 2014), trained on a corpus of 10,876,004 English abstracts of biomedical articles from PubMed with 1,701,632 distinct words (types). The implementation of the above features was accomplished using *Summaryrank*[2], a reference package released for a similar task (R.-C. Chen, Spina, Croft, Sanderson, & Scholer, 2015; L. Yang et al., 2016).

Sentence Ranking

As mentioned above, there might be multiple sentences that answer the query, but only the most suitable should be extracted. For example, while the next two phrases positively answer the question, only the second one is chosen as a claim. *In patients with hypertension, does treatment with ACE inhibitors, compared to placebo, reduce risk of cardiovascular event or improve blood pressure?*

- The vascular pathophysiologic alterations of ISH-a decreased aortic distensibility-can be improved with long-term lisinopril treatment, whereas values deteriorate further in placebo-treated subjects. [PMID: 11336102][assertion= YES]

[2] `https://github.com/rmit-ir/SummaryRank`

- These results, in one of the first studies including subjects with previously untreated ISH only, indicate that lisinopril treatment might favorably influence the cardiovascular risk of ISH. [PMID: 11336102][assertion= YES]

Learning to Rank (LTR) is better suited for such a task since it differs from traditional machine learning techniques; the latter solves it as a classification problem while the aim is an optimal order of the instances in the list (T.-Y. Liu et al., 2009). In our research, we evaluated two of the popular state-of-art learning to rank algorithms,*LambdaRank* and *LambdaMART*. LambdaRank is a succesor of RankNet that only uses the gradient of the costs instead of the model score. LambdaMART benefits from the strengths of MART, Multiple Additive Regression Trees, and LambdaRank by combining regression trees boosting, used in MART with a cost function derived from LambdaRank (Burges, 2010). Our proposed model implements a LambdaMART function, because it outperformed lambdaRank, with training metric NDCG@10. The Normalized Discounted Cumulative Gain (NDCG) is a cumulative measure of the ranking quality truncated at a particular rank level (Järvelin & Kekäläinen, 2000). The model is trained on the generated feature vectors using the RankLib library[1] and the top ranked answer sentence is regarded as the output.

4.4.2 Contradiction Detection

The contradiction detection component is not regarded as a yes/no question answering system, but more as a semantic relation analyzer between two sentences. The system determines whether the input text has an entailment or contradictory relation.

Query Reformulation

This step aims at modifying the PICO-format question into a reduced list of keywords in a declarative form. In our approach, we consider each word in the question as a keyword unless it is a stop word, question word, or the substring ”compared to placebo” is removed as it adds no value when identifying entailment or contradiction. Following that, we apply ClausIE (Del Corro & Gemulla, 2013), an open information extractor, to identify relations and corresponding arguments found in input question.

[1] https://sourceforge.net/p/lemur/wiki/RankLib/

Features

Three features are used to identify the assertion values of claims.

Negation The presence of negation is still the most effective feature of identifying oppositeness. Instead of relying only on the odd count of negation words in the sentence, our proposed model uses NegEx (Chapman, Bridewell, Hanbury, Cooper, & Buchanan, 2001). NegEx takes as input a keyword/concept and a sentence, and uses regular expressions and a predefined trigger word list to decide whether the concept is negated or affirmed. This module iterates three times over each question triple(*left argument,relation,right argument*).

Antonyms The model includes direct and indirect antonyms; for two words *w1* and *w2*,it checks if *w1* is an antonym of *w2* or an antonym of any of its synonyms and vice versa. Instead of comparing raw words, we use lemmas of words for better detection. However, even though the occurrence of antonym pairs in text is a direct and reliable indication of contradiction, it is limited by the low number of antonym pairs in current lexicons. Trying to overcome this limitation, we expand the antonym coverage by using two lexical resources, WordNet (Miller, 1995) and VerbOcean (Chklovski & Pantel, 2004). Below is an example that contains an antonym in *ManConCorpus*:

In women with pre-eclampsia, does treatment with L Arginine, compared to placebo, reduce blood pressure or pre-eclampsia

- L-Arginine load in pregnant women is associated with increased nitric oxide (NO) production and hypotension. [PMID: 10486782 - Assertion value: NO]

In this example, 'reduce' and 'increased' are not direct antonyms like 'good' and 'bad' but are still detected in model. This feature is computed as the count of antonyms per sentence.

Alignment It is also important to include features that model text entailment. Alignment between sentences relies on mapping dependency graphs of two sentences with each other. The algorithm uses SpaCy[2] to generate dependencies, and a built-in similarity score is calculated for each word node in the query related to a similar one in the claim. Finally, the total alignment score is the sum of all output scores.

[2] https://spacy.io/

Classification

A linear support vector machine classifier is used to determine the relation of each input sentence, based on the output feature values. The model implements the classifier using the Scikit library (Pedregosa et al., 2011a).

4.5 Results

4.5.1 Claim Extraction Results

To evaluate the performance of the Learning to Rank framework and the efficiency of the features employed, we conduct two experiments. We first test the model using the first 5 features mentioned in section 4.1 and then we repeat the test after adding the domain-based features covered by the word embedding trained on biomedical articles. For that purpose, we split the *ManConCorpus* into two sets for training and testing purposes. The training set consists of all abstracts with structured format, while the test set includes all unstructured abstracts. After the preprocessing phase, the candidate set has a total of 1212 and 339 sentences for training and testing, respectively. The test set includes 69 answers to only 15 of the 24 queries, while the training set covers all queries with 190 correct claims. Table 4.1 shows the performance results of the claim selection component. The authors in (Alamri & Stevenson, 2015) relied on lexical similarity and a *Z-score* that computes the sentence relevance, with respect to the distribution of similarity scores of other sentences across the dataset. However, While this scoring function contributes to precision, it also affects the recall performance metric. The robustness of our proposed answer detection component relies on the combination of semantic and context features, with an effective ranking algorithm that ranks the sentences according to relevancy, instead of only classifying them as relevant/irrelevant.

	Precision		Recall		F1	
	Ans	Non-ans	Ans	Non-ans	Ans	Non-ans
AlAmri et al.(2015)	0.56	0.92	0.57	0.92	0.56	0.92
Model(GF)	0.92	1	1	0.67	0.96	0.80
Model(GF&DF)	0.94	1	1	0.75	0.96	0.86

Table 4.1: Claim extraction Results. Abbreviations: GF = General features, DF = Domain-based features

4.5.2 Contradiction Detection Results

Performance comparison between models is a non-trivial task, therefore we deploy the same evaluation metrics as in (Alamri, 2016a). Since there is a bias in the distribution of *YES/NO* classes in the corpus, the results are best reported through precision, recall and F1. The baseline performance is measured by annotating all claims with the majority class *YES*. All evaluation results are shown in table 4.2. Our model was able to improve the accuracy of detecting contradictions, namely the *NO* category, and still maintain good results regarding the entailment. The achieved improvement is due to the enhanced negation detection through the NegEx framework, and the inclusion of antonyms.

	Precision		Recall		F1	
	Ent.	Cont.	Ent.	Cont.	Ent.	Cont.
Baseline	0.69	0.0	1.0	0	0.82	0.0
AlAmri et al.(2016a)	0.85	0.80	0.94	0.60	0.89	0.69
Proposed Model	0.95	0.85	0.93	0.89	0.94	0.87

Table 4.2: Contradiction Detection Results

4.6 Conclusions and Future Work

In this paper, we are interested in identifying conflicting findings reported in biomedical literature. We focus on information found in the abstracts as it summarizes all research methodology and conclusion, and conveys important findings without redundancy. It divides the extraction process into two phases, finding the relevant sentences and detecting contradiction. The model combines both semantic and domain-bases features, to enhance the claim detection process. It relies on an SVM classifier that integrates negation, antonyms and alignment scoring to detect conflicting statements. The evaluation results are very promising, specifically in the contradiction detection component, achieving better performance than other systems.

5 Biomedical Textual Inference

The medical literature suffers from disagreements among authors discussing the same topic or treatment. With thousands of articles published daily, there is a need to detect inconsistent and often contradictory findings. Natural language inference (NLI) gained a lot of interest in the past years, however, domain-specific NLI systems are yet to be examined in depth. In this paper, we conduct several experiments on sentence pairs extracted from PubMed abstracts, to infer whether they express entailment, contradiction or neutral meanings. The main focus of this research is to recognize textual entailment in published evidence-based medicine findings. We explore popular NLI models and sentence embeddings, adapted to the biomedical domain. We further investigate improving the inference detection abilities of the models by incorporating traditional machine learning (ML) features with deep learning (DL) architecture. The proposed model serves in capturing biomedical language's representations by combining lexical, contextual and compositional semantics.

This work was originally published as:
Tawfik, N. S., Spruit, M. R. (2019b) Towards Recognition of Textual Entailment in the Biomedical Domain. In *International Conference on Applications of Natural Language to Information Systems* (pp. 368–375). Springer, Cham.

5.1 Introduction

In the last decade, the rate of conducting clinical and medical research has changed dramatically, in terms of both quantity and quality. Subsequently, the number of published results in forms of research papers, clinical trials and textbooks has witnessed a growth spurt. Catillon's synthesis (Catillon, 2017b) estimates that the number of clinical trials has increased from 10 per day in 1975 to 55 and 95 in 1995 and 2015 respectively. In 2017, the PubMed repository contained around 27 million articles, 2 million medical reviews, 500,000 clinical trials and 70,000 systematic reviews. Contribution of medical research is evaluated according to its applicability in the clinical practice and its ability to aid future research in the same field. It is then critical to assess and resonate with published findings specifically when there is more and more evidence on disagreements and contradiction between outcomes (J. P. A. Ioannidis, 2005; Prasad et al., 2012b)

Our work aims to improve the process of evaluating scientific contributions by detecting textual inference between results reported in biomedical abstracts. This paper proposes a model for labeling sentence pairs as entailed, contradictory or neutral. The model relies on linguistic and domain-specific handcrafted features and recent state-of-the-art sentence encoders. The novelty of our approach is the integration of conventional machine learning features with an encoder-based deep neural network.

5.2 Related Work

Textual entailment has been widely studied in recent years, with the availability of SNLI, MultiNLI corpora. However, most models fail to generalize across different NLI benchmarks (Talman & Chatzikyriakidis, 2018), moreover they do not perform accurately on domain-specific datasets. In this section we review textual inference models built specifically for the medical domain. Preclude (Preum, Mondol, Ma, Wang, & Stankovic, 2017a) focuses on extracting conflicts found in health discussions posted in online forums on various health-related topics. The system follows a linguistic rule-based approach to detect inter-advice conflicts. It utilizes MetaMap for semantic clause extraction and tokenization, and then assigns polarity to extracted pairs. More recently, Zadrozny et al. suggested a conceptual framework based on the mathematical sheaf model to highlight conflicting and contradictory criteria in guidelines published by accredited medical institutes. It transforms natural language sentences to formulas with parameters, creates partial order based on common

predicates and builds sheaves on these partial orders (Zadrozny & Garbayo, 2018).
There were few scattered attempts on extracting contradictions from scientific articles avaialable online. Sarafraz et al. (Sarafraz, 2012b), extracted negated molecular events from biomedical literature using a hybrid of machine learning features and semantic rules. Similiarly, De Silve et al. (de Silva, Dou, & Huang, 2017b), extracted inconsistencies found in miRNA research articles. The system extracts relevant triples and scores them according to an appositeness metric suggested by the authors. Alamri et al.(Alamri, 2016b), detected contradictory findings through n-grams, negation, sentiment and directionality. Our previous work combined a ranking model to find the most relevant finding per abstract and detected biomedical contradictions through semantic features and biomedical word embeddings(N. S. Tawfik & Spruit, 2018a).

5.3 Methods

5.3.1 Dataset

In 2016, Alamri et al. published a dataset of contradictory research claims for medical sentence classification and question answering. It is constructed out of 24 systematic reviews on 4 popular cardiovascular disease topics. Medical experts manually mapped each systematic review to a question and extracted corresponding answers from abstracts of articles referenced in the reviews. Only the most relevant sentence is chosen as answer, it is given a *YES* label if it positively answers the question or *NO* label otherwise. More details on the annotation criteria, process and the corpus statistics can be found in (Alamri, 2016b). While the dataset is annotated by experts, its structure is not aligned with the language inference task. For that reason, we reconstruct the corpus by combining claims to build a pairwise-sentence corpus to match conventional NLI datasets. We first fetch the PubMed article ids of all 259 abstracts included in the dataset, and extract the first sentence of each abstract. The first sentence in an abstract often describe the research objective. We enrich the corpus by adding extracted sentences and assigning them with the label *NEUTRAL*. Our choice of objective sentence to fill as neutral is based on the general observation of neutral sentences across different NLI benchmarks where they are usually constructed by adding a purpose clause (Gururangan et al., 2018). Given the unique set of medical questions denoted Q where each question is related to only one systematic review and multiple abstracts. For each q_i that belongs to Q, we assumed the following hypotheses while labeling the sentence pairs as entailed, contradictory or neutral:

- *$claim_2$* entails *$claim_1$* if *asr_2=YES* AND *asr_1=YES*
- *$claim_2$* contradicts *$claim_1$* if *asr_2=YES* AND *asr_1=NO*
- *$claim_2$* contradicts *$claim_1$* if *asr_2=NO* AND *asr_1=YES*
- *$claim_2$* is neutral to *$claim_1$* if *asr_2=YES* AND *asr_1=NEUTRAL*
- *$claim_2$* is neutral to *$claim_1$* if *asr_2=NEUTRAL* AND *asr_1=YES*

Where *asr* denotes the assertion value of each sentence with three possible values *YES, NO, NEUTRAL. Claims* refer to the question answer extracted from abstracts. It is important to note that for formulating the above guidelines, a definition of 'entailment' and 'contradiction' is needed. Therefore, we follow the original corpus interpretation of contradiction as "Two texts, T1 and T2, are said to contradict when, for a given fact F, information inferred about F from T1 is unlikely to be true at the same time as information about F inferred from T2". The final dataset consisted of 2135 sentence pairs with 1080, 608 and 447 entailment, contradiction and neutral class instances respectively.

5.3.2 Machine Learning

Human Engineered Features

The model has a total of 20 traditional NLP features divided into 3 main categories. The main selection criteria of features was to capture context, lexical and semantic representations of text with a limited and optimized feature set.
String-based features This sub category includes *editDist, LevSim, CosSim, JacSim* to represent shortest/longest edit distance, Levenshtein similarity, Cosine similarity and jaccard similarity respectively. In addition, we calculate 4 variations of length measures between the two sentences: *LenMax, LenMin, LenAbs, LenAvg*
Contradiction-based features Negation is still a robust measure of appositeness, we define 4 features to detect negation in sentences. *NegationBin* as a binary feature, *NegOverlap* as the jaccard similarity of negated words only, *AntScore* as a score between the count of antonyms found between sentences. To expand the antonyms coverage we use both WordNet and VerbOcean lexicons, and also *ModOverlap* as the similarity between modal words found in both input. In addition to the above set we also try to detect the outcome polarity through Subjectivity and sentiment (*SubjScore, SentLabel*) using the NLTK

sentiment analyzer. Moreover, the results sentence of scientific articles are often accompanied by a ”change clause” that affects the final output (Niu, Zhu, Li, & Hirst, 2005). The key is to detect whether changes occurring in both sentences are bad, good or neutral. To measure the final pairwise polarity, we include more features such as *PolarityBin* as a binary feature set to 1 when both sentences share the same polarity and 0 otherwise, and *ChangePolarity* that scores each sentence according to a predefined list of change keywords labelled good(+ve score values) or bad(-ve score values).
Context-based features To include domain knowledge we add *EntityOverlap* that computes the similarity between medical UMLS concepts identified by MetaMap[1]. We also rely on word embeddings to capture context. Our hypothesis is that models trained on domain knowledge would generate vector representation capable of learning conceptual meaning of the domain. We compute *EmbedSim* as the cosine similarity between the two embedding vectors and the *EmbedAvg* as the similarity between embedding average for each sentence pooling of all word embeddings. The word embeddings are extracted using FastText model pre-trained on the PubMed Central open access subset [2] . We add the Word Mover’s Distance *WMDSim* as measure of similarity between both sentences.

Classification

We experiment with different classification algorithms available in the Scikit-learn toolkit. The experiments include Support Vector machine, Linear regression model, Random Tree, Gradient boost and Naive Bayes.

5.3.3 Deep learning

Sentence Embeddings

Text embedding are considered a key element in various NLP tasks. Popular word embeddings such as Word2Vec and GloVe outperform existing models that rely on co-occurrence counts because of their ability to better represent distributional semantics. To encode sentences with one of the prior models, a simple average of their corresponding word embeddings would yield strong results. Nonetheless, during the last two years we witnessed a rise of different supervised and unsupervised approaches towards learning representations of sequences of words, such as sentences or paragraphs. They are able to identify

[1] `https://metamap.nlm.nih.gov/`

[2] `https://github.com/lucylw/pubmed_central_fasttext_pretrained`

the order of words within a sentence and hence capture more context. The developed sentence representations extend the success of earlier word vectors with interesting results and increasing potential in different tasks. We focus our research on the two of the most popular sentence encoding schemes InferSent and Universal Sentence Encoder. We argue that fine tuning these models and leveraging transfer learning could possibly lead to a good performance in a domain-specific settings. Both chosen encoders were trained partially or fully on textual inference data which fits perfectly with our task.

InferSent is a sentence encoder proposed by Facebook(Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017). Its main advantage over other models is its supervised training over SNLI, a large text inference dataset manually annotated. The original model [3] is trained on 570k human-generated English sentence-pairs with a bi-directional Long Short Term Memory (BiLSTM) encoder.
*Universal Sentence Encoder (USE)*was developed by Google (Cer et al., 2018). It has two variations, the first is a transformer-based encoder which yields high-accuracy at the cost of high complexity and extra computational resources. The second model uses a deep averaging network that averages word embeddings and serve as input to a deep neural network. In our model, we deploy the transformer architecture as it was proven to yield better results in several NLP tasks. The universal sentence encoder [4] training data contains supervised and unsupervised sources such as Wikipedia articles, news, discussion forms, dialogues and question/answers pairs. It is also partially augmented with instances from the SNLI corpus.

Deep Learning Network

Our DL model follows a siamese-like architecture where the first set of layers are parallel duplicates and share same weights. For merging the two inputs, we concatenate the element-wise difference and then multiply both vectors. Following that, there are multiple intermediate dense layers. The nodes are directly connected to the nodes in the next layer and use rectified linear activation (ReLU) function. Given the small dataset size, we introduce a dropout layer with a dropout rate of 0.3. Finally, the prediction layer with 3 nodes predicts the probability of each of the inference classes, and a softmax activation function. We adopt an exponentially decaying learning rate, and an l2

[3] Pre-trained model for InferSent available at `https://github.com/facebookresearch/InferSent`

[4] Pre-trained model for USE available at `https://tfhub.dev/google/universal-sentence-encoder-large/3`

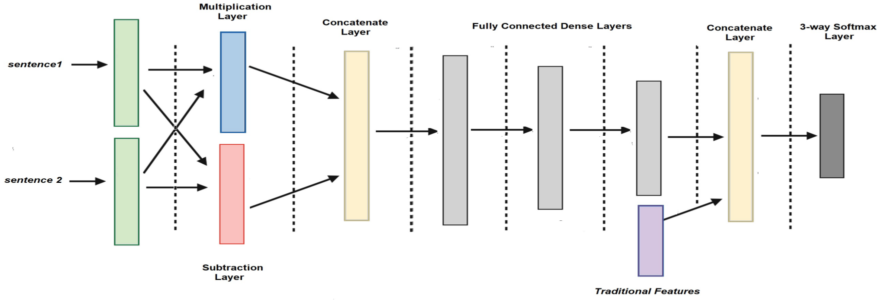


Figure 5.1: The feature-assisted neural network architecture.

regularization weight of 0.01.

5.3.4 A feature-Assisted Neural Network Architecture Model

With the small size of the dataset, traditional features demonstrate good performance in comparison with the neural network models. This, along with more evidence on the usefulness of combining traditional features in deep learning architecture (R.-C. Chen, Yulianti, Sanderson, & Bruce Croo, 2017; Sequiera et al., 2017), encouraged us to build a hybrid model. An essential dilemma for building the feature-assisted model is how to incorporate engineered features to sentence embeddings inputs. Directly appending the traditional ML features to the encoded representations generated from InferSent or USE would not influence the performance. In that scenario, the features' effect on the classification decision would almost be nonexistent given the large size of sentence encoding vector versus the feature vector size of 21 values. Figure 7.1 gives an overview of the final feature-assisted framework we propose.

5.4 Results and Evaluation

All the following results are calculated as the average results of standard cross-validation with 10 folds. The results reported for the machine learning approach are the output of the best two classifiers: Random Forest (RF) and extreme gradient boosting (XGBoost). It is generally observed that XGBoost almost always achieves higher accuracy than RF. Table 5.1 shows the results details of the model, The baseline performance is 50.56% based on the majority classifier output. We note, that the ML experiments were not meant for direct comparison with the DL model. The conducted evaluations serve at choosing the best feature combination that could further boost the DL network. As for the deep learning algorithms, we ran multiple experiments while varying the number of hidden layers and the corresponding number of nodes. Adding more layers test our model capacity, in other terms, with small number of layers the model may struggle to fit the data. On the other hand, over-scaling the network size leads to great results on training data and performs poorly on the test data. Our experiments show that there was a minimal overfitting effect with increasing the number of layers, however, there was no added accuracy. Deep Learning experiments' results are shown in table 5.2. In all cases, InferSent encoder outperforms USE encoder with approximately 8%. This finding is consistent with previous published findings (Q. Chen, Kim, Wilbur, Du, & Lu, 2018). Both encoders are considered universal and should represent sentences efficiently given the amount of data they are trained on. The performance difference between the two encoders could be attributed to the difference in the embedding vector dimension (512 vs 4096) and the nature of inference data *InferSent* is trained on. We added the traditional features to the best performing model with 3 layers and a number of nodes decreasing by 50% with each hidden layer that is deeper in the neural network. No remarkable acheivment were noticed in the *USE* encoder case(only 0.6% difference). However, the hybrid model achieves the best result with an average accuracy of 96.21% and a minimum of 94.32% when combined with the *InferSent* encoder. Even with a limited dataset, the results suggest that the machine learning features and deep learning models are complementary. Their combination in an end-to-end model can enhance the learning process and improve the predictions on unseen data.

Table 5.1: Machine learning features with Random Forest and XGBoost classifiers based on 10-fold cross validation. Reported numbers correspond to average accuracy and standard deviation

Feature set	**Random Forest**	**XGBoost**
context-based	53.26% (+/- 1.80%)	49.16 % (+/- 2.67%)
contradiction-based	67.81% (+/- 1.28%)	69.49% (+/- 1.77%)
context + string	61.01% (+/- 1.97%)	64.91% (+/- 2.98%)
all features	72.30% (+/- 2.32%)	**76.94% (+/- 1.24%)**

Table 5.2: Deep Learning performance results on 10-fold cross validation with respect to the number of hidden layers in the DNN architecture. Reported numbers corresponds to average accuracy and standard deviation

Hidden layers	**Hidden Units**	**USE** *(Dim.:512)*	**InferSent** *(Dim:4096)*
1 layers	512	72.56% (+/- 1.14%)	89.88% (+/- 3.91%)
3 layers	512,256,128	82.27% (+/- 1.63%)	**93.95% (+/- 1.39%)**
3 layers	512,256,64	83.17% (+/- 2.20%)	93.86% (+/- 1.48%)
5 layers	512,256,256,128,128	83.68% (+/- 1.50%)	92.24% (+/- 0.79%)
3 layers	512,256,128,64,64	83.68% (+/- 1.50%)	93.18% (+/- 1.73%)

5.5 Conclusion

We attempt to detect medical text inference from published scientific articles. Various experiments have been applied in different scenarios including ML features and DL network built on top of sentence encoders. Our proposed hybrid architecture is the optimal configuration in terms of size and number of hidden layers. The final results are promising, however, the model must be re-evaluated on a larger corpus to generalize its effect. We could enhance the sentence encoder power by re-training them on domain-specific sources such as research articles and clinical notes. We also believe that a feature ablation analysis over a bigger range of features could potentially select a better boosting vector for assisting the neural network.

6 | Biomedical Text Representations

Text representations are one of the main inputs to various Natural Language Processing (NLP) methods. Given the fast developmental pace of new sentence embedding methods, we argue that there is a need for a unified methodology to assess these different techniques in the biomedical domain. This work introduces a comprehensive evaluation of novel methods across ten medical classification tasks. The tasks cover a variety of BioNLP problems such as semantic similarity, question answering, citation sentiment analysis and others with binary and multi-class datasets. Our goal is to assess the transferability of different sentence representation schemes to the medical and clinical domain. Our analysis shows that embeddings based on Language Models which account for the context-dependent nature of words, usually outperform others in terms of performance. Nonetheless, there is no single embedding model that perfectly represents biomedical and clinical texts with consistent performance across all tasks. This illustrates the need for a more suitable bio-encoder. Our MedSentEval source code, pre-trained embeddings and examples have been made available on GitHub.

This work was originally published as:

Tawfik, N. S., & Spruit, M. R. (2020b) Evaluating Sentence Representations for Biomedical Text: Methods and Experimental Results. *Journal of Biomedical Informatics*, 103396.

6.1 Introduction

In the past few years, neural network-based distributional representations of text such as word embeddings have been shown highly effective in solving multiple NLP problems. There are many different types of word embeddings (Schnabel, Labutov, Mimno, & Joachims, 2015). However, they all have the same purpose: to generate low-dimensional vector representations of words. They can encode important syntactic properties of words efficiently, and are able to capture semantic similarity among words as mathematical similarities between their vectors. Similarly, sentence embeddings are numerical representations of sentences, which are often derived from word embeddings. Nonetheless, the last two years witnessed a rise of different supervised and unsupervised approaches towards learning representations of sequences of words, such as sentences or paragraphs. They can identify the order of words within a sentence and hence capture more context. The developed sentence representations extend the success of earlier word vector-based approaches with interesting results and increasing potential across different tasks.

The progress in machine learning has given scientists the unprecedented opportunity to extract valuable information from biomedical data. With the increasing availability of unstructured textual data in the biomedical domain in forms of clinical trials, research articles, electronic health records, and patient-authored texts, the use of text mining techniques is becoming increasingly more important. The importance of word embeddings in Biomedical Natural Language Processing (BioNLP) becomes evident by looking at the number of recent researches in the field. These embeddings have been commonly leveraged as feature input for several BioNLP tasks. Word-level embeddings have been studied extensively in the biomedical domain (Z. Chen, He, Liu, & Bian, 2018; Chiu, Crichton, Korhonen, & Pyysalo, 2016; Y. Wang, Liu, Afzal, et al., 2018). On the other hand, the analysis of sentence-level representations has been much more limited to a few scattered works (Q. Chen, Peng, & Lu, 2018; Hao, Liu, Wu, & Lv, 2018) and there is a lack of a full analysis of embedding techniques on common grounds.

Motivated by (Conneau & Kiela, 2018; Perone, Silveira, & Paula, 2018), in their efforts to evaluate sentence representations in a fair and structured approach, this paper aims at replicating their evaluations in domain-specific settings. More specifically, we assess the ability of existing sentence representation techniques to capture the rich and complex semantics of clinical sentences. We focus on what are arguably the state-of-art techniques in embedding sentences known for achieving high performance in general NLP tasks. Throughout our analysis, we test and compare several sentence embedding methods trained on general, medical and clinical data. Our evaluations include multiple classification problems related to the clinical and biomedical domain spanning different linguistic tasks. We discuss the strengths and weaknesses of the different techniques in encoding domain-specific aspects of clinical sentences. To our knowledge, no similar evaluation exists for the biomedical domain. This paper is organized as follows: In Section 2, we give a background of word and sentence embeddings. In Section 3, we provide details of all ten tasks included in our evaluations. We also describe the experimental settings of the sentence embedding models implemented. In Sections 4 to 6, we illustrate the results obtained and draw corresponding conclusions accordingly.

6.2 Background

Words representations are inspired by the concept of distributional semantic models that hypothesize that word meanings could be inferred by the company they keep (Mackin, 1978). While the concept is old, its recent popularity could be traced back to the work of Bengio et al. on natural language modeling through a neural probabilistic model (Bengio et al., 2003). Among the first approaches to embed words based on neural networks is the Word2vec algorithm proposed by Mikolov et al. (Mikolov, Chen, Corrado, & Dean, 2013). The model is a shallow, three-layered neural network that uses unsupervised learning to determine the semantic and syntactic meanings of a word based on adjacent words denoted as context. It offers two variations: Continuous Bag of Words (CBOW) and Skip-gram. The first learns the representations by predicting the target word based on its context words while skip-gram inverts contexts and targets, and tries to predict each context word from its target word, rather than predicting the target word itself. Global Vector word representations (GloVe) (Pennington, Socher, & Manning, 2014) is another prominent method that enabled efficient unsupervised training of dense word representations and straightforward integration into NLP tasks. GloVe is a count-based model, as opposed to Word2Vec that is considered a predictive model. Count-based models utilize the word distribution statistics of the corpus effectively. It constructs a global word-word co-occurrence matrix and applies matrix factorization to learn lower dimension embeddings, where each row is some vector representation for each word. FastText is a recent addition to prediction-based models, proposed by Facebook, for learning word embeddings over large datasets. Its architecture is similar to the skip-gram model, but it includes a significant improvement that accounts for the morphological properties of words through the learning process (Bojanowski, Grave, Joulin, & Mikolov, 2017). Embeddings are produced by combining the n-gram embeddings for all the n-grams characters in a word. The main advantage over prior techniques is its ability to predict out-of-vocabulary (OOV) words as a result of its sub-word representations. The above models are static, context independent and do not account for polysemy. In other words, the model outputs only one vector for each word regardless of the word position in the sentence, or the context in which it appears (Yaghoobzadeh & Schütze, 2016).

On the other hand, embeddings based on Language Models (LM) dynamically change so they can discriminate among different meanings of a word. Language modeling serves as an unsupervised pre-training stage, where learning is independent of the main task, on a large unlabeled or differently-labeled text corpus. It can generate the next word in a sentence with knowledge of previous words. These resulting embeddings are the internal states of deep neural networks in a monolingual or a bilingual language modeling setting. Among the first attempts to generate context-sensitive representations is Context2vec (Melamud, Goldberger, & Dagan, 2016). The model represents the context of a target word by extracting the output embedding of a multi-layer perceptron built on top of a bi-directional Long short-term memory (LSTM) language model. Other examples include Embeddings from Language Models (ELMo) (M. E. Peters et al., 2018), Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, Google, & Language, 2018), Universal Language Model Fine-tuning (ULMFiT) (Howard & Ruder, 2018) and the Pooled Contextualized Embeddings from Flair toolkit (Akbik, Blythe, & Vollgraf, 2018).

ELMo Vectors are also computed on top of two-layer bidirectional Language Models (biLMs) with character convolution. Using CNNs, each vector is built upon the characters that compose the underlying words. BERT is different from ELMo primarily because it targets a different training objective; it uses masked language modeling instead of traditional LM. It overcomes ELMo limitations by including left and right contexts simultaneously when representing words. BERT replaces words in a sentence randomly and inserts a "masked" token. The transformer generates predictions for the masked words based on left and right unmasked neighbors. FLAIR contextualized embeddings are word-level embeddings that were shown effective in the sequence labeling task. Input sentences are modeled as distributions over sequences of characters to a bidirectional character-level language model. The neural model is pre-trained on large unlabeled corpora, and internal character states are used to compute the output word embeddings.

To obtain sentence embeddings on top of word embeddings, a simple Bag-

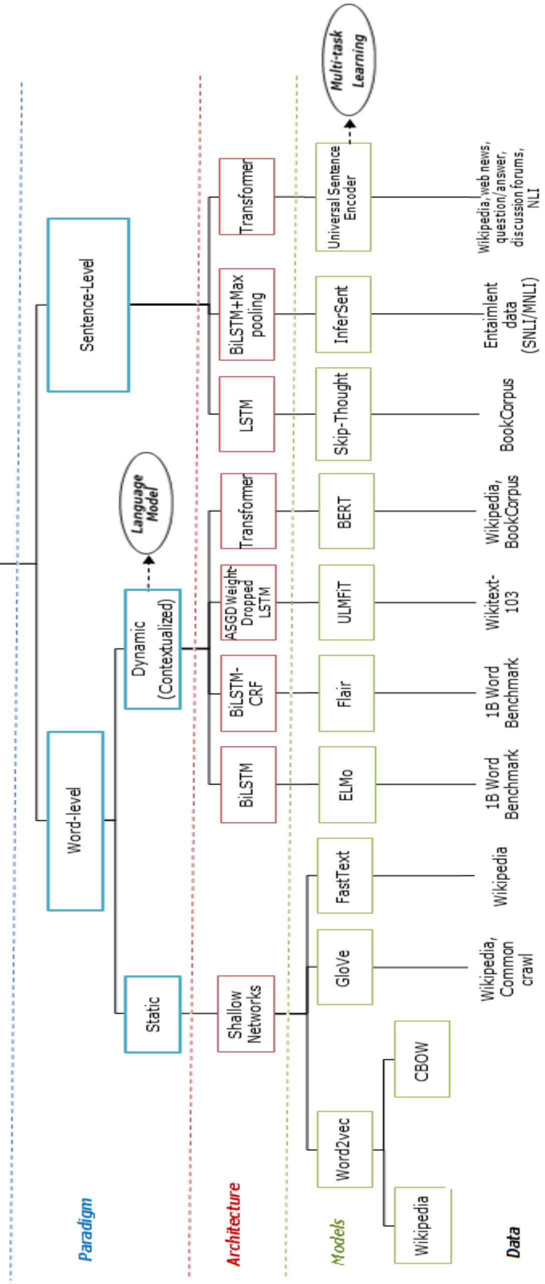


Figure 6.1: Classification of different embedding models

of-Words (BoW) inspired method could be applied by computing the mean of the vector embeddings for the words in a sentence. Alternatively, more advanced and sophisticated methods such as Sent2Vec and Smooth Inverse Frequency (SIF) could be employed. In the Sent2Vec paradigm, a sentence embedding is defined as the average of the source word embeddings of its constituent words (Pagliardini, Gupta, & Jaggi, 2018). The method is furthermore augmented by learning source embeddings for unigrams and n-grams of words present in each sentence, and averaging the n-gram embeddings along with the words. In contrast, SIF adds a weighting function to word embeddings, which down-weights common words (Arora, Liang, & Ma, 2017). However, despite the improved performance achieved using sophisticated methods, the main limitation of conventional embeddings at the word-level is the negligence of the overall sentence structure. Nevertheless, some scholars argue that regardless of the potential information loss, word embeddings are still able to represent sentence meanings efficiently (Adi, Kermany, Belinkov, Lavi, & Goldberg, 2017; H. Li, Caragea, Li, & Caragea, 2018).

Alternatively, there have been efforts to generate dedicated sentence embeddings through unsupervised training. The Skip-Thought model extends the original skip-gram algorithm from words to sentences (R. Kiros et al., 2015). It predicts neighbour sentences or phrases for a given sentence using a recurrent neural network. It follows an encoder-decoder model where first a sentence is encoded into a vector through a Gated Recurrent Units (GRU) or LSTM architecture, and that representation is decoded into surrounding text. It adopts a vocabulary expansion scheme that makes use of pre-trained embeddings to learn embeddings of new non-encountered words.

In contrast, Facebook introduces inferSent (Conneau et al., 2017), a supervised learning methodology for sentence encoding. Their work provides solutions to two critical concerns: the best training task and network architecture to obtain a universal sentence representation model. Their findings indicate that detecting natural language inference is the most suitable for transfer learning to other NLP tasks. This is attributed to the semantic nature of the task and the availability of a very large corpus such as the Stanford Natural Language Inference (SNLI) that consists of 570k humanly generated English sentence pairs, manually labeled. Moreover, they experiment with seven different architectures including standard recurrent encoders with either LSTM or GRU, concatenation of last hidden states of forward and backward GRU, Bidirectional LSTMs (BiLSTM) with either mean or max pooling, self-attentive networks, and hierarchical convolutional networks. They conclude that the BiLSTM with the max-pooling operation performs best on both SNLI and transfer tasks. Google also published a sentence encoder known as Universal

Sentence Embeddings (USE) (Cer et al., 2018). It is referred to as "universal" since, in theory, it is supposed to encode general properties of sentences given the large size of datasets it is trained on. The multi-task learning encoder uses several annotated and unannotated datasets for training. It has two variants of the encoding architectures. The Transformer model is designed for higher accuracy, but the encoding requires more memory and computational time. The Deep Averaging Network (DAN) model, on the other hand, is designed for speed and efficiency, and some accuracy is compromised. When integrated with any downstream task, USE should be able to represent sentences efficiently without the need for any domain-specific knowledge. This is a great advantage when limited training resources are available for specific tasks.

We include GloVe, FastText, ELMo, BERT, Flair, Infersent, and USE embeddings in our evaluations as we believe that they successfully represent all different techniques previously discussed.

6.3 Methods

In this paper, we propose *MedSentEval*[1], an embedding evaluation toolkit designed for the medical domain. It can compute, evaluate, and classify pre-trained sentence embeddings for several BioNLP tasks. The proposed toolkit heavily makes use of SentEval[2], a general evaluation protocols toolkit. This section describes the included tasks and gives further details on the pre-trained models supported in our toolkit and the evaluation procedures for each task.

6.3.1 Evaluation tasks

One of the main challenges in the biomedical NLP domain is the availability of benchmark corpora for evaluation. The creation of a dataset faces many barriers such as privacy issues for patient data protection due to the sensitive nature of data, the inefficacy of using crowd-sourcing platforms to annotate data and the need to rely on domain experts which endures more costs. Despite the limitations mentioned above, there have been efforts in creating datasets. However, they are relatively small in size and mostly focus on information extraction. In this paper, we gathered BioNLP datasets that are suitable for classification problems and cover a variety of NLP tasks, including binary and multi-class classification. Below, we provide a brief description of each dataset

[1] `https://github.com/nstawfik/MedSentEval`

[2] `https://github.com/facebookresearch/SentEval`

grouped by types of tasks. Table 6.1 summarizes the details of each of the datasets used in our experiments and provide examples.

Textual Entailment (TE) TE is an important task in the NLP domain. Given two snippets of text, Text (T) and Hypothesis (H), the TE recognition determines if the meaning of H can be inferred from that of T (Dagan et al., 2013a). The medical natural language inference benchmark dataset *MedNLI* is a source of biomedical TE data derived from clinical notes (“A Natural Language Inference Dataset For The Clinical Domain”, 2018; Romanov & Shivade, 2018). Its creation process is similar to the creation of the gold-standard SNLI dataset with adaptation to the clinical domain. Expert annotators were presented with 4,638 premises extracted from the MIMIC-III database (Johnson et al., 2016) and were asked to write three hypotheses with true, false, and neutral descriptions of each premise. The final dataset comprises 14,049 sentence pairs divided into 11,232, 1,395 and 1,422 for training, development and testing, respectively.

Recognizing Question Entailment *RQE*, tackles the problem of finding duplicate questions by labeling questions based on their similarity (Ben Abacha & Demner-Fushman, 2016; “The Medical Question Entailment Data”, 2016). Extending the former textual entailment definition, the authors define question entailment as ”Question A entails Question B if every answer to B is also a correct answer to A exactly or partially.” The *RQE* dataset is specifically designed to find the most similar frequently asked question (FAQ) to a given question. The training set was constructed from the questions provided by family doctors on the National Library of Medicine (NLM) platform resulting in 8,588 question pairs where 54.2% are positive pairs. For the test set, two sources of questions were used: validated questions from the NLM collections and FAQs retrieved from the National Institutes of Health (NIH) website. The test set corpus includes 302 pairs of questions, with 42.7% pairs positively labeled.

Sentence classification In recent years, there has been a substantial increase in the number of scientific publications in the biomedical domain with valuable evidence-based medicine guidelines. Consequently, many NLP methods were deployed to automate or semi-automate the analysis of large medical literature databases such as PubMed. Both datasets included in this category use randomized controlled trials (RCT) as a source of data. The *PICO* dataset is curated from abstracts of RCTs available in the PubMed repository (Jin & Szolovits, 2018; “PubMed PICO Element Detection Dataset”, 2018). It maps each abstract sentence into the known Participants, Intervention, Comparison, and Outcome (PICO) elements (Richardson, Wilson, Nishikawa, & Hayward, 1995). Moreover, the authors extended the original PICO framework and added three additional categories: aim, method, and results. The annotated data include 24,668 abstracts; each sentence was assigned to a category according to a predefined keyword list compiled manually. The final dataset contains approximately 257K, 31K, and 30K sentences for training, testing, and validation. The *PUBMED20K* corpus is designed for sequential sentence classification of RCTs textual data. Abstracts' sentences are labeled according to their role in the abstract into background, objective, method, result, or conclusion (Dernoncourt & Lee, 2017; "PubMed 200k RCT Dataset", 2017). The data collection process was limited to randomized controlled trial abstracts with a structured format. The dataset is large in size with around 180K sentences for training, 30K sentences for validation, and another 30K sentences for testing with a total of 20,000 abstracts.

Sentiment analysis Reproducibility is very common in biomedical research where many studies try to replicate earlier work. Scientists express their opinions in many different ways, specifically when citing other studies. The citation sentiment analysis corpus *CitationSA* ("Citation Sentiment Analysis Dataset", 2015; Xu et al., 2015) is the first of its kind in the biomedical domain. It includes the discussion section of 285 randomly selected clinical trial abstracts. The annotation scheme did not consider the correctness of the published claims and polarity was assigned on the citation level according to context and not at the sentence level. Two medical annotators labeled sentences according to the former scheme, and a third annotator was involved in case of disagreement. The total number of citation sentences included in the dataset is 4,182 citations. It is also important to include a dataset that represents patient opinions on health-related topics. Patient authored texts usually mix medical terminologies with informal language. *VaccineSA* is a collection of English tweets that includes HPV vaccination keywords (Du, Xu, Song, Liu, & Tao, 2017; "HPV Vaccination's Tweets Dataset", 2017). The tweets were classified according to their content and polarity into positive, negative, neutral, and unrelated. The negative category is further divided based on the negative concern such as safety, efficacy, cost, resistant, or other. The dataset originally contained 3,984 tweets, however, when we recollected the tweets, only 1,853 were available.

Question answering This task is a longstanding problem extensively studied in the past years and is currently gaining interest in the biomedical domain. The BioASQ challenge ("Biomedical Semantic Question Answers", 2018; Tsatsaronis et al., 2015) targets different stages of the question answering process, ranging from the retrieval of relevant concepts and articles to the generation of natural language answers. For the classification task, our interest is in the second phase of task B where BioASQ released questions from benchmark datasets created by a group of biomedical experts. The questions are accompanied by text snippets extracted from relevant PubMed and PMC articles. Four question types are included in the challenge: yes/no, factoid, list, and summary questions, we experiment with the yes/no category only. The BioASQ 6b-task dataset includes 612 question-answer pairs for training and 130 pairs for testing.

We also add the Bio-Contradiction *BioC* dataset to the evaluation. Although it was originally built to detect contradictions among published findings, the dataset structure is also suitable for the question answering task (Alamri & Stevenson, 2016b; "A Corpus of Contradictory Research Claims from Cardiovascular Research Abstracts", 2016). It is organized into 24 PICO questions related to cardiovascular disease generated from systematic reviews. Two annotators were asked separately to curate all research abstracts of studies referenced in the systematic review and extract one sentence per abstract that answers the question. Sentences are labeled *YES* if they positively answer the question and NO otherwise. The corpus has a total of 259 question-answer pairs, out of which 180 are labeled *YES* and 79 labeled *NO*.

Semantic Text Similarity (STS) Different than the former classification tasks, the goal of this task is to measure the relatedness of two sentences and compare it with a human-labeled similarity score. Clinical Semantic Textual Similarity *ClinicalSTS* was published as part of the shared task at the 2018 BioCreative/OHNLP challenge ("Clinical Semantic Textual Similarity Dataset", 2018; Y. Wang, Liu, Rastegar-Mojarad, et al., 2018). This challenge was organized to investigate the STS problem in the clinical domain following the lead of the original SemEval STS shared tasks. The dataset is a randomly annotated subset of the MedSTS dataset that consists of a total of 174,629 sentences. The dataset was collected from patients records at the Mayo Clinic's clinical data warehouse. Three surface lexical similarities were employed to find candidate pairs: Ratcliff/Obershelp pattern matching algorithm, cosine similarity, and Levenshtein distance.

More details on the original MedSTS dataset construction could be found

Table 6.1: Evaluation datasets description and examples

Dataset	Task	Source	Example	Label
MedNLI	Textual Entailment	Patient records	*H1:During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB* *P2:The patient is on room air*	Contradiction
RQE	Question Entailment	Doctor questions	*Q1: What should I do with this patient whose biopsy report shows carcinoma in situ of the vulva?* *Q2: What to do with this patient, biopsy shows carcinoma in situ of the vulva?*	True
PUBMED20K	Sentence Classification	Medical articles	*Text:Transient intraocular pressure elevation and cataract progression occurred .*	Background
PICO	Sentence Classification	Medical articles	*Text: Classes included CRC survivors and people with CVD .*	Intervention
PatientSA	Sentiment Analysis	Patient tweets	*Text: Don't forget to also vaccinate your sons. It is potentially even more important. #HPV #vaccineswork*	Positive
CitationSA	Sentiment Analysis	Medical articles	*Text: Patrek et al [C] examined 13 factors influencing fluid drainage.*	Neutral
BioASQ	Question Answering	Medical articles	*Q:Is osteocrin expressed exclusively in the bone?* *A:Evolution of Osteocrin as an activity-regulated factor in the primate brain.*	No
BioC	Question Answering	Medical articles	*Q:In women with pre-eclampsia, is mutation in renin-angiotensin gene associated with pre-eclampsia?* *A:The variants(A->C) of 1166 polymorphism site of AT1RG predisposes increased risk of PIH.*	Yes
C-STS	Semantic Similarity	Patient records	S1: Use information was down loaded from the patient's PAP device and reviewed with the patient. S2:I discussed the indications, contraindications and side effects of doxycycline with the patient.	0.5
BIOSSES	Semantic Similarity	Medical articles	*S1: The oncogenic activity of mutant Kras appears dependent on functional Craf.* *S2: Oncogenic KRAS mutations are common in cancer.*	1

in (Y. Wang, Afzal, et al., 2018). For ClinicalSTS, the sentence pairs were annotated independently by two clinical experts who scored each pair based on their semantic equivalence. Scores ranged from 0 to 5, where 0 denotes complete dissimilarity between sentences. The final similarity value was set to the average of both annotators' scores. The dataset includes 1,068 sentence pairs with 70% (750 sentence pairs) and 30% (318 sentence pairs) for training and testing, respectively. All included sentences are de-identified sentences as all protected health information (PHI) was removed through a frequency filtering approach followed by a manual check. The second corpus, the biomedical sentence similarity estimation *BIOSSES* corpus ("Biomedical Semantic Similarity Estimation System", 2017; Sogancioglu, Öztürk, & Özgür, 2017) comprises 100 sentence pairs. All sentences are extracted from the biomedical summarization track of the Text Analysis Conference (TAC). The subset includes sentences from biomedical articles with a citation mention to a reference article. Similar to *ClinicalSTS*, the dataset creators followed the SemEval guidelines for annotations with five experts giving 0 to 4 score values to sentence pairs to indicate no relation (0) or equivalent(4).

6.3.2 Embedding Methods

GloVe We use the pre-trained embeddings consisting of 2.2 million vocabulary words available at `https://nlp.stanford.edu/projects/glove/` which were trained on the Common Crawl (840B tokens) dataset. The authors in (Newman-Griffis, Lai, & Fosler-Lussier, 2017) trained GloVe on the 2016 PubMed baseline and made them publicly available at `https://slate.cse.ohio-state.edu/BMASS/`.

FastText General embeddings were obtained from `https://fasttext.cc/docs/en/english-vectors.html` also trained on the Common Crawl corpus resulting in 2 million word vectors. For the domain-specific pre-trained model, we used the embeddings provided at `https://github.com/lucylw/pubmed_central_fasttext_pretrained`.

ELMo We use the original 5.5B configuration, as recommended by the authors, trained on Wikipedia and news crawl data. To further investigate the embedding size effect on performance, we added the small model trained on the 1 Billion Word Benchmark. Moreover, we download their biomedical domain contributed model trained on PubMed. ELMo embeddings are computed after concatenating all three layers of the ELMo. All models were downloaded from `https://allennlp.org/ELMo` and implemented through the AllenNLP python toolkit (Gardner et al., 2018).

BERT We evaluated both base and large base models provided at `https://`

github.com/google-research/bert. We also take advantage of two newly released BERT models: BioBERT (Lee et al., 2019) trained on the PubMed abstracts with a vocabulary size of 4.5B words and SciBERT (Beltagy, Cohan, & Lo, 2019) trained on scientific articles from the biomedical and computer sciences domains with 2.5B and 0.6B word count, respectively. The pre-trained weights of the BioBERT model (version 1.0/ PubMed 200K) are available at https://github.com/naver/biobert-pretrained and for the SciBERT model at https://github.com/allenai/scibert. We use the original Google BERT GitHub repository to encode sentences; it originally provides fine-tuning scripts for the pre-trained model in an end-to-end fashion. It additionally describes how to obtain fixed contextual embeddings of each input token generated from the hidden layers of the pre-trained model. Following the steps to obtain the embeddings, we only use the final hidden layer of the transformer (layer value set to -1). The maximum sequence length was set to 128 with a batch size of 32 as per the authors' recommendation.

Flair The authors recommend using both forward and backward Flair embeddings. We chose the mixed model as it is trained over a diverse corpus including web, Wikipedia, and Subtitles for the English language. They also provide pre-trained embeddings over 5% of PubMed abstracts until 2015. All models are downloaded from and implemented through the official Flair repository https://github.com/zalandoresearch/flair.

InferSent The authors provide two versions of the model, one based on GloVe embeddings and another based on FastText embeddings. We experiment with both available at https://github.com/facebookresearch/InferSent. InferSent is based on supervised learning from natural inference data. For this model, we train our own models using the Medical natural language inference (MedNLI) dataset using the biomedical Glove and FastText embeddings mentioned earlier.

USE In our evaluation, we use the transformer-based architecture of the USE encoder as it was proven to yield better results. Training data consisted of supervised and unsupervised sources such as Wikipedia articles, news, discussion forums, dialogues and question/answer pairs. USE was implemented through its TF hub module available at https://tfhub.dev/google/universal-sentence-encoder-large/3.

6.3.3 Experimental setup

A fundamental dilemma is how to compare different models that vary between vanilla embeddings, contextualized embeddings, or dedicated sentence encoders. The performance differences between models could be attributed to many reasons such as better pre-trained word embeddings, the different architecture, the different objective, or the normalization layer (Wieting & Kiela, 2019). SentEval was created to overcome the limitation of the non-standard evaluation of embedding methods across research, especially when a considerable number of embeddings techniques have surfaced in the last years. SentEval's evaluation strategy relied on simple classification models. The choice behind the basic classifiers was to assess how these representations fare without the use of complex models like recurrent neural networks (RNNs). As in the original SentEval tooklit, we opted for the same models as it allows us to observe the benefits of different embedding models in representing and predicting linguistic medical information of the input text. Our choice was also supported by several studies following the SentEval methodology in evaluating new models (J. Kiros & Chan, 2018; Perone et al., 2018; Reimers & Gurevych, 2019; Wieting & Kiela, 2019). We follow the same guidelines in our experimental settings. For tasks that require classification, we experiment with both logistic regression and multi-layer perceptron (MLP) on top of the generated sentence representations. The MLP consists of a single hidden layer of 50 neurons using Adam optimizer and a batch size of 64 with a Sigmoid non-linearity function.

Extracting word embeddings from vanilla models such as GloVe and FastText is a straightforward process. On the other hand, the contextualized models offer two paradigms to deploy their pre-trained models to adapt to the target task: feature extraction and fine-tuning. The former is similar to feature-based models where pre-trained weights are kept frozen whereas, in the latter, the weights are trained further on the new task. In a recent paper, Peters et al. compare the effectiveness of both adaptation methods. Their results show that BERT generally performed better in the feature-extraction mode while the opposite is exact for ELMo. Their evaluation also proves that the performance depends on the similarity of the pre-training and target tasks. The feature-extraction mode aligns with our strategy to standardize the comparison criteria across all evaluation experiments. Not only because the use of embeddings as features is the only possible method for other models but also because the fine-tuning process differs from BERT to ELMo. In MedSentEval, ELMo features are calculated for each token by concatenating all three layers weights of the model. For BERT features, we take the hidden states of the final hidden layer of the transformer model. To generate sentence vectors from word embeddings, we apply the Mean of Word Embeddings (MOWE) technique (White, Togneri, Liu, Bennamoun, & Ben, 2015) on both static and contextualized embeddings.

Table 6.2: Experimental settings for each evaluation task

Task	**Classes**	**Classification**	**Validation**	**Performance Metrics**
MedNLI	3	LR/MLP	Standard validation	Accuracy
RQE	2	LR/MLP	Standard validation	Accuracy
PUBMED20K	5	LR/MLP	Standard validation	Accuracy
PICO	8	LR/MLP	Standard validation	Accuracy
PatientSA	8	LR/MLP	Nested cross-validation	Accuracy
CitationSA	3	LR/MLP	Nested cross-validation	Accuracy
BioASQ	2	LR/MLP	Cross-validation	Accuracy/F1
BioC	2	LR/MLP	Nested cross-validation	Accuracy/F1
C-STS	[0-5]	-	Cosine similarity	Pearson/Spearman correlation
BIOSSES	[0-4]	-	Cosine similarity	Pearson/Spearman correlation

In tasks with dual inputs as in textual entailment tasks, their combined embedding vector is built as $(u, v, \mid u - v \mid, u * v)$, which is a concatenation of the premise and hypothesis vectors and their respective absolute difference and Hadamard product. Tasks evaluation criteria are consistent across tasks with minor variations as their input/output formats, and types are different. For example, for semantic similarity tasks, we only need to calculate the cosine similarity between the input's embedding and compare it to the expert-labeled score through Pearson and Spearman correlations.

All experiments were carried out using a single GPU with 12 GB RAM. However the toolkit also provides optional support of CPU only machines through scikit-learn for the logistic regression. Since this might trigger memory issues with some datasets such as *PICO* and *PubMed20K*, we only recommend this for small sized datasets. Table 6.2 highlights the experimental settings for each task and the performance metrics used for evaluation. Our analysis do not include the time factor when conducting the comparison since both the feature extraction and classification phases do not exceed 2 hours for all tasks with the exception of BERT and ELMo models when run over big datasets namely *PICO* and *PubMed20K*. We note that this does not include the time needed to generate the pre-trained weights as it may cost more time to train some embedding models, such as InferSent, Bert or ELMo.

6.4 Results

In Table 6.3, shows results obtained from the included embedding schemes across all 10 tasks included in *MedSentEval*. The reported results are based on the logistic regression classifier as it consistently achieved better results than MLP on most transfer tasks and specifically on small size datasets. ELMo takes the lead with the original 5.5B model excelling in 5 out of 10 tasks in the general embeddings category. The PubMed version is also dominating with 4 tasks in the embeddings acquired from biomedical training data. The BERT algorithm comes next with the best performance of 2 and 3 tasks for base and BioBert models respectively. Moreover, BERT embeddings are often the second best performing on many tasks with minimal accuracy difference from ELMo which did not exceed 1%. Figure 6.2 and 6.3 illustrate a comparison between each method with general and domain-specific training. Additionally, we investigate whether there is any correlation between independent model factors and perceived performance. Driven by several research questions, we analyzed the results of the conducted evaluation.

Static versus Context embeddings Comparing the representation models within each category separately, we find that context-dependent models capture more information than regular static embeddings. The only exception to that rule was the BioASQ dataset, where Glove and FastText achieve better results than ELMo, BERT and Flair in the biomedical domain and BERT and ELMo only in the general domain. This exception is not indicative as we additionally observe that tasks in the question-answering category, in general, are the least influenced by the different techniques since the classifier tends to overfit to the majority class in most models.

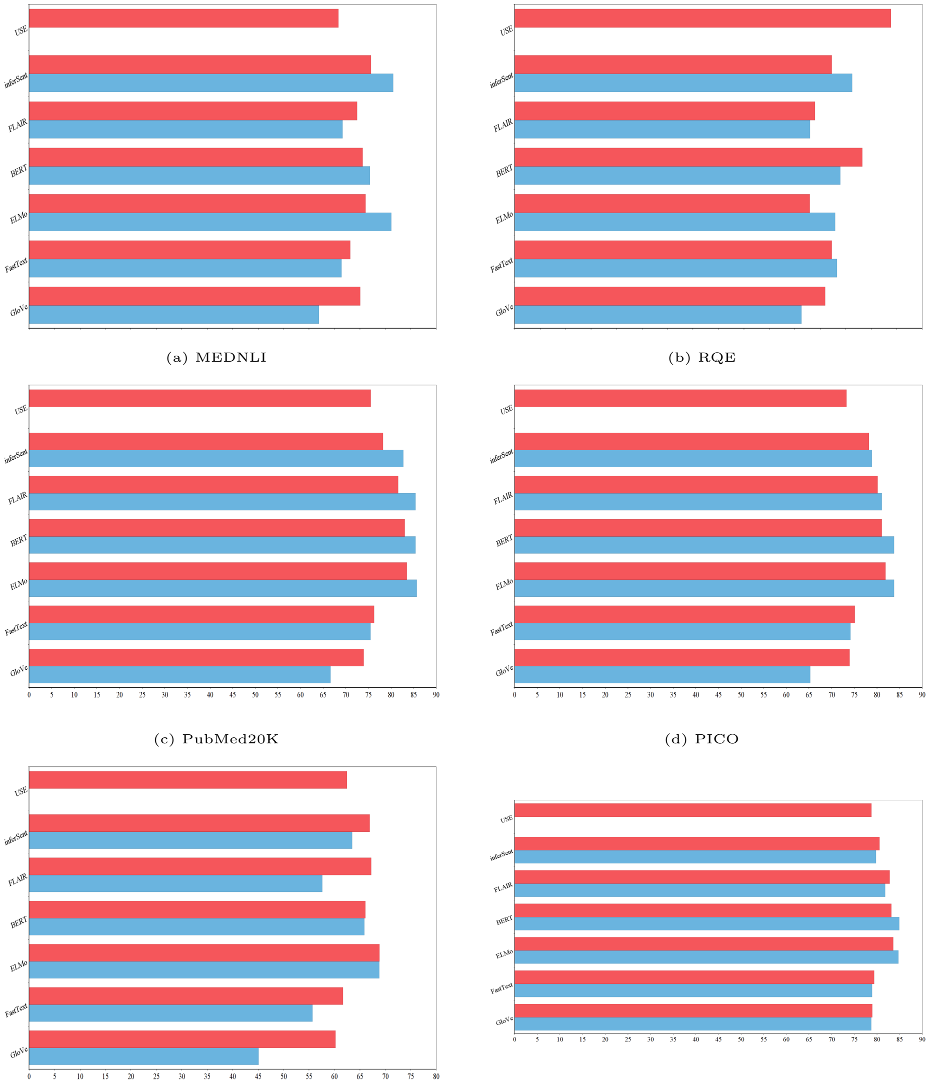
General versus Domain embeddings Apart from GloVe, FastText, and Flair, comparing each general embedding model to its biomedical peer, the latter always outperforms the former. In the case of Flair, the medical embeddings are worst in performance or do not provide a significant gain over the general model. As mentioned, ELMo and BERT are the best-suited models in both the general and biomedical categories to represent medical text.

Word-based versus Sentence-encoders embeddings Under the assumption that sentence-level encodings better capture the content of medical text since it takes into account the word order within the sentence, we observed that the best results are generally obtained through averaging word embeddings. The reason for this result may be related to the fact that much of the word order information is captured in general natural language word order statistics (Adi et al., 2017). This observation is true for all tasks except for the language inference task. However, we believe that this is due to the similarity of the task's data and the training data of the sentence encoders. The inferSent supervised model is trained on the SNLI and MedNLI datasets for the general and biomedical embeddings categories, respectively. While the Universal Sentence Encoder has multi-type data including questions and entailment pairs, among others.

Embedding Dimension versus Embedding model As we employ the averaging scheme to calculate the sentence embedding, the size of the embedding vector is equal to the original word vector size. While FLAIR and InferSent have the biggest embedding dimension of 4096, their performance is inferior to other models with a much smaller embedding vector. On the other hand, in models like ELMo and BERT, where we experiment with different versions of the same model with different embedding dimensions, we notice that increasing the embedding vector size is related to a performance gain. This finding is expected as the more dimensions a word vector has, the more semantic information can be preserved in the resulting representation. This explains the poor performance of GloVe and FastText biomedical embeddings. The embedding size of the pre-trained biomedical model are 200 and 100 for GloVe, and FastText respectively, while the general embedding dimension is 300.

Table 6.3: Logistic regression performance on all tasks included in *MedSentEval.* For all tasks we report accuracies except for the BioC/BioASQ we additionally report F1. And for BIOSSES and ClincalSTS, we report Pearson/Spearman correlations between the cosine distance of both sentences and the similarity score given by the domain expert. Underlined values indicate the best overall result, while values in **bold** indicate best performing over each category.

Tasks	**Dim**	**MedNLI**	**RQE**	**PubMed20K**	**PICO**	**Vac.SA**	**Cit.SA**	**BioC**	**BioASQ**	**C-STS**	**BIOSSES**
General											
GloVE$_{840B}$	300	65.05	60.93	73.95	73.95	60.20	78.97	69.89/82.20	69.23/81.82	0.27/0.48	0.25/0.39
FastText$_{crawl}$	300	63.08	62.25	76.23	75.09	61.67	79.39	69.51/82.01	69.23/81.82	0.54/0.56	0.43/0.50
ELMo$_{small}$	768	62.03	58.28	80.26	78.68	62.22	81.73	74.12/83.48	66.15/78.35	0.54/0.50	0.33/0.35
ELMo$_{Org5.5B}$	3072	66.10	57.92	**83.51**	**81.89**	**68.82**	**83.60**	**81.44/87.81**	67.02/78.80	0.57/0.49	0.29/0.32
Bert$_{baseC}$	768	63.99	65.56	81.91	80.73	64.54	83.18	74.89/84.66	66.92/79.62	**0.69/0.57**	0.48/0.50
Bert$_{largeUc}$	1024	65.54	68.21	83.04	81.08	66.07	82.80	74.51/84.09	69.23/81.30	0.66/0.51	**0.56/0.58**
Flair$_{news}$	4096	61.12	56.62	81.58	80.15	67.21	82.79	72.20/83.30	66.92/79.62	0.54/0.47	0.31/0.30
Flair$_{mix}$	4096	64.42	58.94	81.26	79.60	65.62	81.54	70.28/82.39	**70.00/82.03**	0.52/0.53	0.40/0.50
InferSent$_1$	4096	**67.16**	58.22	74.70	61.42	66.91	81.86	79.94/97.18	70.00/81.86	0.57/0.54	0.32/0.42
InferSent$_2$	4096	63.99	62.25	78.24	78.24	64.28	80.55	69.51/82.01	68.46/81.28	0.54/0.57	0.43/0.48
USE	512	60.76	**73.84**	75.50	73.26	62.46	78.76	69.50/82.01	69.23/81.82	0.64/0.56	0.45/0.48
Biomedical											
Glove$_{PubMed}$	200	56.96	56.29	66.61	65.26	45.08	78.71	69.50/82.01	**69.23/81.82**	0.08/0.42	0.05/0.17
FastText$_{PubMed}$	100	61.39	63.25	75.47	74.16	55.67	78.92	69.50/82.01	**69.23/81.82**	0.28/0.56	0.56/0.60
ELMo$_{PubMed}$	3072	71.18	62.91	**85.71**	83.76	**68.79**	84.73	**83.00/88.34**	66.92/80.00	0.61/0.54	**0.74/0.70**
BioBERT	768	66.95	63.91	85.35	83.69	65.50	83.91	74.51/84.21	68.46/81.28	**0.69/0.54**	0.64/0.62
SciBERT	768	66.24	63.91	85.44	**83.77**	65.86	**84.93**	82.25/88.41	67.69/80.00	0.67/0.59	0.56/0.60
Flair$_{PubMed}$	2300	61.60	57.95	82.73	81.08	57.61	81.81	72.61/83.47	66.92/79.81	0.40/0.48	0.35/0.47
InferSent$_{MedNLI}$	4096	**71.52**	**66.23**	79.54	78.86	63.47	79.80	71.05/82.69	68.85 /81.55	0.54/0.53	0.35/0.41



(a) MEDNLI

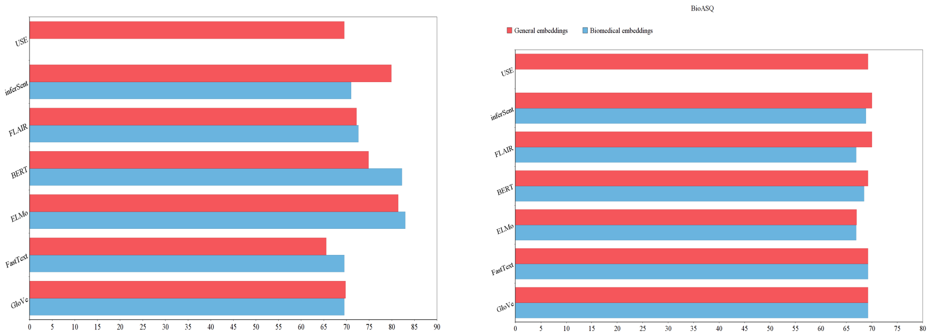
(b) RQE

(c) PubMed20K

(d) PICO

(e) VaccineSA

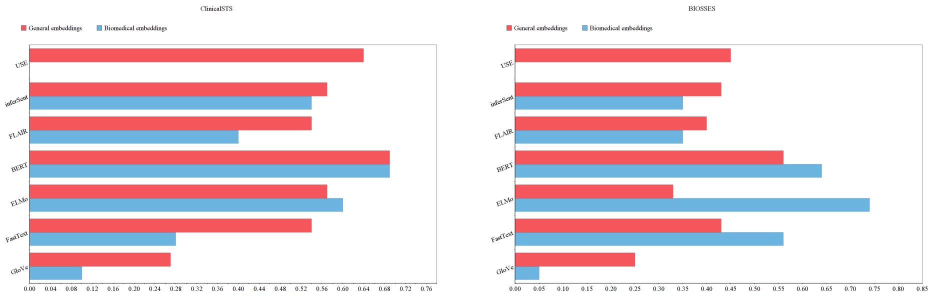
(f) CitationSA



(g) BioC

(h) BioASQ

Figure 6.2: Accuracy values for the logistic regression classifier across tasks included in *MedSentEval*.



(a) C-STS

(b) BIOSSES

Figure 6.3: Results for the Semantic Text Similarity tasks. Values shown are the Pearson correlation coefficients for the test sets.

6.4.1 Qualitative Analysis

Besides the quantitative results mentioned above, we also test the effectiveness of the different biomedical models on several simple, but non-trivial, examples beyond the extrinsic tasks. In this section, we shed light on intrinsic characteristics that may explain some of the performance variances. Building on our previous findings that domain-based embeddings are better suited for the medical and clinical NLP tasks, we limit the qualitative analysis to the biomedical embedding models. The diagnostic data used throughout this analysis have been manually selected to fit the test purpose. However, they do not represent the language distribution as a whole. Our main goal was to provide insight into what models are capturing, what are their strengths and weaknesses through adversarial examples.

General Knowledge We first attempt to investigate the robustness of the generated numerical representations to reflect common sense (Z. Yang, Zhu, & Chen, 2019). The test consists of removing non-important tokens, such as stop words, and calculate the similarity between the original and the shortened sentences. For this test, we collect ten sentences, one from each dataset, while varying the number of stop words per chosen sentences. Table 6.4 shows an example of the sentences before and after removal, and the corresponding cosine distances computed by the six biomedical models. The models are ranked according to the descending order of the similarity values. The higher the value, the better the ability of the model to assign similar embeddings for both sentences. Consider the example in table 6.4, most models still retains a similarity of 0.93 or above ,except for GloVe, even after removing 13 stop words from a single sentence. More examples are available in Appendix A. The results demonstrate that the InferSent model is the most insensitive regarding the removal of non-important words and, surprisingly, ELMo's performance is not consistent. In many cases, the similarity decreases by 10%, ranking below FastText in all 10 examples. The same applies to Flair embeddings; this suggests that preprocessing the text before embedding with these models could give higher priority to informative words and might yields better results in downstream tasks.

Concept Identity The second test measures to what extent the sentence representation encodes the identities of entities within it. In the medical and clinical language, referring to concepts using their abbreviations is a common practice and frequently found in sentences extracted from both patient records and scientific articles. Retaining the concept identity, whether it is referred to in full or in abbreviated form, is crucial and demonstrates the models' capacities pertinent to language understanding.

We collected five examples with abbreviated concepts in the premises from the MedNLI dataset. We compared the NLI predicted labels given the original premise and after expanding the abbreviations. It is clear from observing the examples in table 6.4 and Appendix A that BERT is able to relate the premise to the hypothesis more often when using the full-form leading to the correct inference. On the other hand, ELMo has zero gain after the expansion process. While the number of cases is relatively small to generalize the results, based on this intuition, we suggest expanding and normalizing abbreviations and acronyms in the dataset before using BERT.

Domain Knowledge The third test assesses the model's ability to capture significant semantic meanings of the input text. In the context of our domain-based evaluation, the semantic significance mainly refers to the medical information within the sentence (Weng & Szolovits, 2019). For this test, we consider 6 records that hold indirect information relating independent medical concepts. In the example shown in table 6.4, most models fail to relate *chest pain* to *Angina.* Consequently they interpret these to be two unrelated entities and labels the instance pair as neutral. This pattern is consistent in most models across different relation categories such as *Disease-Symptom*, *Drug-Disease*, *Drug-Drug classes.* While empirical results show that all the representations encode certain amount of information, our finding advocates for integrating external knowledge to compensate for the lack of medical background of the models. This could be achieved by adding external knowledge sources such as the Unified Medical Language System (UMLS) to include semantic types and relationships.

6.5 Discussion

This paper inspects sentence representations for biomedical text by analyzing seven popular embedding schemes. It is important to recall that the primary purpose of this study is not to outperform existing state-of-the-art methods for the reported BioNLP tasks. We are seeking to evaluate and validate different embedding techniques that will enable further in-depth investigations and improvement of text representations for the biomedical domain. When comparing with baseline performance introduced in datasets' original papers, if applicable, the results obtained is close to or outperforms baseline values. This is mainly due to the use of simple classification algorithms like the MLP or LR classifiers.

Therefore, the performance achieved through the toolkit still has room for improvement by fine tuning the hyperparameters of the classification models

Table 6.4: Qualitative Analysis

Test Objective	**Example**	**Predicted**			**Expected**
General Knowledge	*Original: Our findings suggest an association between the DD genotype of the ACE gene and early-onset but not later-onset pre-eclampsia which may give a partial explanation for the higher recurrence risk with early-onset pre-eclampsia .* *Modified: Our findings suggest association DD genotype ACE gene early-onset later-onset pre-eclampsia may give partial explanation higher recurrence risk early-onset pre-eclampsia .*	FastText	0.97		~1
		InferSent	0.97		
		BERT	0.96		
		ElMo	0.87		
		Flair	0.84		
		Glove	0.8		
Concept Identity	*Premise: Reports lack of appetite but no n/v.* *Expanded premise: Reports lack of appetite but no nausea and vomiting.* *Hypothesis: the patient denies nausea and vomiting.*		Pre	Post	E
		GloVe	N	C	
		FastText	N	E	
		ElMo	C	C	
		BERT	C	E	
		Flair	N	C	
		InferSent	C	E	
Medical Knowledge	*Premise: No chestpain or fevers .* *Hypothesis: Patient has no angina*	GloVe	C		E
		FastText	E		
		ElMo	N		
		BERT	N		
		Flair	C		
		InferSent	E		

or deploying other classifiers. It was reported that employing complex models such as Convolutional Neural Networks (CNNs) is more effective for text classification tasks (Kim, 2014). Similarly, using the end-to-end BERT model could lead to a non-trivial accuracy gain for several tasks such as MedNLI (Alsentzer et al., 2019; N. Tawfik & Spruit, 2019). However, the effect of classification algorithms on the performance and the analysis of different approaches to adapt the pre-trained representations (M. Peters, Ruder, & Smith, 2019) are out of the scope of this paper.

The benchmark gives insights on the sentence embedding quality through downstream tasks with four extrinsic tasks (Textual Entailment, Sentence classification, Sentiment analysis, and Question answering). Additionally, we follow a case-based reasoning approach to provide a qualitative analysis of the learned representations. We found relatively modest correlations between the quantitative and qualitative results. In the case of ELMo, for example, the intrinsic evaluation results fail extrinsic performance. This also matches previous evaluations outside the medical domain (Hollenstein, de la Torre, Langer, & Zhang, 2019). Finally, both evaluation types could benefit from domain experts' perspectives. It is hard to draw conclusions or rank models from best to worst as there is no single sentence embedding scheme that consistently performs well on all of the ten tasks. As with all classification problems, specific approaches are better suited to some datasets than others, this is also consistent with the "no free lunch theorem" (Wolpert & Macready, 1997). However, our experimental results unveil a number of important observations:

- Sentence embeddings computed as the mean of word embeddings are still effective in capturing the sentence semantics and yield competitive results to dedicated sentence encoders.

- There is no correlation between the embedding dimension and the performance across different models.

- In almost all cases, neural embeddings generated from hidden states of a deep learning model are able to capture more semantics than word embeddings computed from count or prediction based models.

- Contextualized word embeddings with a language model objective, i.e. ELMo and BERT, usually outperform other encoding schemes.

- While InferSent is better suited for textual entailment, given the type of data it is trained on, its good performance does not generalize over other tasks.

- A proper balance and variation in the training resources, when compared to training solely on domain data, can lead to more efficient results such as the case of BioBert and SciBERT.

The results show that most models still need to resume training on domain-related and task-specific data. And that, to date, producing a single universal embedding model that generalizes well to other tasks requires more investigations and evaluations. Given the superiority of ELMo and BERT over other models, we particularly recommend integrating language models with neural embeddings as a promising direction of research. Incorporating medical knowledge in the learning process of the models is also expected to enhance their performance. Another alternative for improving the results is to combine two or more embedding techniques in a single classification model. Adopting such a method could offer specific domain background, when adding concept embeddings for example, to acquire the best of each technique.

6.6 Conclusion

In this paper, we presented *MedSentEval*, a new toolkit for evaluating state-of-the-art sentence embedding methods for NLP classification problems. Through our evaluations, we assessed the transferability of these embeddings to biomedical domain tasks. Our research aimed to build on the work done by Conneau et al. (Conneau & Kiela, 2018) and adapt it to fit medical and clinical text corpora. We also integrated extra embedding techniques not available in the original toolkit such as ELMo and BERT. We hope that our in-depth evaluations, along with the toolkit, will benefit the BioNLP community in selecting suitable embeddings for different application tasks. While only downstream tasks are used to evaluate the overall quality of sentence representation models, we also note the need to incorporate probing tasks as in *SentEval*. A future extension of our current work will include support for more embeddings schemes trained on different domain data types such as patient records, nurse notes and full-text articles PubMed central combined. Moreover, adding more tasks for each category, when available, could further improve our understanding and generalization of the findings.

7 Clinical Textual Inference

This article describes the participation of the *UU_TAILS* team in the 2019 MEDIQA challenge intended to improve domain-specific models in medical and clinical NLP. The challenge consists of 3 tasks: medical language inference (NLI), recognizing textual entailment (RQE) and question answering (QA). Our team participated in tasks 1 and 2 and our best runs achieved a performance accuracy of 0.852 and 0.584 respectively for the test sets. The models proposed for task 1 relied on BERT embeddings and different ensemble techniques. For the RQE task, we trained a traditional multilayer perceptron network based on embeddings generated by the universal sentence encoder.

This work was originally published as:
Tawfik, N. S., Spruit, M. R. (2019) UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain. In *Proceedings of the 18th BioNLP workshop and Shared Task* (pp. 493–499).

7.1 Introduction

Detecting semantic relations between sentence pairs is a long-standing challenge for computational semantics. Given two snippets of text: Premise P and Hypothesis H, textual entailment recognition determines if the meaning of H can be inferred from that of P (Dagan et al., 2013a). The significance of modeling text inference is evident since it evaluates the capability of Natural language Processing (NLP) to grasp meaning and interprets the linguistic variability of the language. Natural language inference (NLI) tasks, also known as Recognizing Textual Entailment (RTE) require a deep understanding of the semantic similarity between the hypothesis and the premise. Moreover, they overlap with other linguistic problems such as question answering and semantic text similarity. The recent years witnessed regular organization of shared tasks targeting the RTE/NLI task, which consequently led to advances in the field. More complex models were developed that rely on deep neural networks, this was feasible with the availability of large amounts of annotated datasets such as SNLI and MultiNLI (Bowman, Angeli, Potts, & Manning, 2015; Williams, Nangia, & Bowman, 2018). However, most models fail to generalize across different NLI benchmarks (Talman & Chatzikyriakidis, 2018). Additionally, they do not perform accurately on domain-specific datasets. This is specifically true in the medical and clinical domain. Compared to open domain data, the language used to describe biomedical events is usually complex, rich in clinical semantics and contains conceptual overlap. And hence, it is difficult to adapt any of the former models directly.

The MEDIQA challenge (Ben Abacha, Shivade, & Demner-Fushman, 2019a) addresses the above limitations through its three proposed tasks. The first task aims at identifying inference relations between clinical sentence pairs and introduces the medical natural language inference benchmark dataset *MedNLI* (Romanov & Shivade, 2018). Its creation process is similar to the creation of the gold-standard SNLI dataset with adaptation to the clinical domain. Expert annotators were presented 4,638 premises extracted from the MIMIC-III database (Johnson et al., 2016) and were asked to write three hypotheses with a true, false and neutral description of the premise. The final dataset comprises 14,049 sentence pairs divided into 11,232, 1,395 and 1,422 for training, development and testing respectively. An additional test batch was provided by the challenge organizers with 405 unlabelled instances.

Similarly, the second task, Recognizing Question Entailment (RQE), tackles the problem of finding duplicate questions by labeling questions based on their similarity (Ben Abacha & Demner-Fushman, 2016). Extending the earlier NLI definition, the authors define question entailment as ”Question A entails Question B if every answer to B is also a correct answer to A exactly or partially”. The dataset is specifically designed to find the most similar frequently asked question (FAQ) to a given question. The training set was constructed from the questions provided by family doctors on the National Library of Medicine (NLM) platform resulting in 8,588 question pairs where 54.2% are positive pairs. For validation, two sources of questions were used: validated questions from the NLM collections and FAQs retrieved from the National Institutes of Health (NIH) website. The validation set has 302 pairs of questions with 42.7% pairs positively labelled. The test set for the challenge was balanced and comprised of 230 question pairs.

The rest of the paper is organized as follows: Section 2 briefly discusses related work. We limit our summary to textual inference research in the biomedical domain only. In Section 3, we describe our proposed model and the implementation details for both tasks. In Section 4, we show the experiment results of our proposed models. Finally, we conclude our analysis of the challenge, as well as some additional discussions of the future directions in Section 5.

7.2 Related Work

In (Ben Abacha & Demner-Fushman, 2016), the authors introduce a baseline model for the RQE dataset. The feature-based model relies on negation, medical concepts overlap and lexical similarity measures to detect entailment among medical question pairs. Romanov and Shivade conducted multiple experiments on the MedNLI dataset to evaluate the transferability of existing methods in adapting to clinical RTE tasks (Romanov & Shivade, 2018). The best performing was the bidirectional LSTM encoder of the inferSent. Their findings also showed that transfer learning over the larger SNLI set did not improve the results. In a previous work, we tried to model textual entailment found in biomedical literature by restructuring an existing YES/NO question-answering dataset extracted from PubMed (S. Tawfik & R. Spruit, 2019). The newly formed dataset aligned with standard NLI datasets format. Further on, we combined hand-crafted features with the inferSent model to detect inference.

To the best of our knowledge, other than the work previously mentioned, there has been minimal research conducted directly on the textual entailment task in the biomedical domain. Below, we summarize scattered attempts to extract contradictions and conflicting statements found in medical documents. Sarafraz et al. (Sarafraz, 2012b), extracted negated molecular events from biomedical literature using a hybrid of machine learning features and semantic rules. Similiarly, De Silve et al. (de Silva et al., 2017b), extracted inconsistencies found in miRNA research articles. The system extracts relevant triples and scores them according to an appositeness metric suggested by the authors. Alamri et al. (Alamri, 2016b), introduced a dataset of 259 contradictory claims that answer 10 medical questions related to cardiovascular diseases. Their proposed model relied on n-grams, negation, sentiment and directionality features while in (N. S. Tawfik & Spruit, 2018a), the authors exploited semantic features and biomedical word embeddings to detect contradictions using the same dataset. Zadrozny et al. (Zadrozny & Garbayo, 2018) suggested a conceptual framework based on the mathematical sheaf model to highlight conflicting and contradictory criteria in guidelines published by accredited medical institutes. It transforms natural language sentences to formulas with parameters, creates partial order based on common predicates and builds sheaves on these partial orders.

7.3 Exploratory Embedding Analysis

With the fast developmental pace of text embedding methods, there is a lack of unified methodology to assess these different techniques in the biomedical domain. We attempted to conduct a comprehensive evaluation of different text representations for both tasks, prior to submission of round 2 of the challenge. We use the *MedSentEval*[1] toolkit, a python-based toolkit that supports different embedding techniques including traditional word embeddings like GloVe and FastText, contextualized embeddings like Embeddings from Language Models (ELMO) and Bidirectional Encoder Representations from Transformers (BERT) and dedicated sentence encoders such as inferSent and Universal Sentence Encoder (USE). To evaluate the sentence representations fairly, we adopt a straightforward method that extracts embeddings from different techniques and feeds them to a logistic regression classifier. Our analysis showed that for the NLI task, embeddings from the inferSent model achieved the best performance. This is not surprising, and aligns with the results reported by the benchmark creator (Romanov & Shivade, 2018). Moreover, we notice that embeddings acquired from language models such as ELMO and BERT, were the second best performing with minimal accuracy difference. For the *RQE* task, the transformer encoder of the USE model outperformed all other methods by a clear margin followed by inferSent trained with GloVe

[1] `https://github.com/nstawfik/MedSentEval`

embeddings. This might be contributed to the multi-type training data employed by USE with questions and entailment sentence pairs among others. As observed in the General Language Understanding Evaluation (GLUE) benchmark dataset, BERT-based models are currently the state-of-the art models for the NLI task. Accordingly, we have tried to further investigate the performance of BERT in the biomedical NLI domain. We also employed USE and inferSent sentence embeddings for task 2.

Bidirectional Encoder Representations from Transformers BERT is a neural model developed by Google, that makes heavy use of language representation models designed to pre-train deep bidirectional representations (Devlin et al., 2018). It is trained in an unsupervised manner over an enormous amount of publicly available plain text data. Language Modeling (LM) serves as an unsupervised pre-training stage that can generate the next word in a sentence with knowledge of previous words in a sentence. BERT is different from other LM-based models because it targets a different training objective, it uses masked language modeling instead of traditional LM. It replaces words in a sentence randomly and inserts a ”masked” token. The transformer generates predictions for the masked words by jointly conditioning on both left and right context in all layers.

Universal Sentence Encoder USE is referred to as ”universal” since, in theory, it is supposed to encode general properties of sentences given the large size of datasets it is trained on (Cer et al., 2018). The multi-task learning encoder uses several annotated and unannotated datasets for training. Training data consisted of supervised and unsupervised sources such as Wikipedia articles, news, discussion forums, dialogues and question/answers pairs. It has two variants of the encoding architectures; The transformer model is designed for higher accuracy, but the encoding requires more memory and computational time. The Deep Averaging Network (DAN) model on the other hand is designed for speed and efficiency, and some accuracy is compromised. When integrated in any downstream task, USE should be able to represent sentences efficiently without the need for any domain specific knowledge. This is a great advantage when limited training resources are available for specific tasks.

7.4 Methods

7.4.1 Task 1: Natural Language Inference (NLI)

Experimental Settings We take advantage of two newly released BERT models trained on different biomedical data. The following models were initial-

Hyperparameter	**Value**
`Learning rate`	3e-5, 2e-5, 5e-5
`Sequence length`	64, 128
`Number of Epochs`	3
`Batch Size`	8, 16

Table 7.1: Hyperparameters values for training BERT models

ized from the original "bert-base-uncased" setting pre-trained with 12 transformer layers, hidden unit size of d=768, 12 attention heads and 110M parameters.

- SciBERT[2] trained on a random sample of 1.14M scientific articles available in the semantic scholar repository. The training data consists of full-text papers from the biomedical and computer sciences domain with a 2.5B and 0.6B word count, respectively (Beltagy et al., 2019).

- ClinicalBERT[3] trained on approximately 2M clinical records. The training data consists of intensive care notes distributed among 15 types available in the MIMIC database (Alsentzer et al., 2019).

We combined both training and evaluation records to form a new training set of 12627 sentence pairs. The original test set was used for evaluation and development. We experimented with all models in pytorch, using the HuggingFace[4] re-implementation of the original BERT python package. We convert the SciBERT models to make it compatible with PyTorch. We use the fine-tuning script to train the model on the MEDNLI dataset in an end-to-end fashion. We trained a total of 30 models with variations of the model configuration. All models with accuracy less than 0.786 on development data were discarded. The threshold value was set to the best accuracy achieved for the MedNLI dataset as reported in the paper. Table 7.1 list the hyperparameters for this set of experiments, the values for other parameters were kept the same as the original BERT model.

[2] The pre-trained weights for for the SciBERT model are available at `https://github.com/allenai/scibert`

[3] The pre-trained weights for the ClinicalBERT model are available at `https://github.com/EmilyAlsentzer/clinicalBERT`

[4] `https://github.com/huggingface/pytorch-pretrained-BERT`

BERT Ensemble Model

Rather than using only a single model for predictions, ensemble techniques can be considered as a useful method to boost the overall performance. A key factor in ensembling is how to blend the results. We experimented with different systems in terms of size and fusion technique in order to increase performance accuracy:

- Drop-out Averaging: All BERT models are added into the candidate ensemble set. Iteratively, we randomly drop one model at a time. With each dropout, we test the ability of the new ensemble set to improve the overall performance by calculating the ensemble's accuracy for the development set by averaging the output probabilities for each class. The process has been repeated until no improvements were observed and the best performing set is chosen as the final ensemble set.
- Stacking BERT 1: A meta learner trained on the predictions generated from all base models and optimally combine them to form the final decision. We train three classifiers, by using five-fold cross validation, including a K-Nearest Neighbor (KNN), a linear Support Vector Machine (SVM) and Naive Bayesian (NB). The classifiers were implemented through the scikit-learn library [5]and we also apply the grid search method for parameter tuning (Pedregosa et al., 2011b).
- Stacking BERT 2: We create a second level ensemble stacking. In this level, we train a logistic regression classifier on top of the combined predictions generated from the first level stacking stacking BERT phase.

7.4.2 Task 2: Recognizing Question Entailment (RQE)

Experimental Settings We use the transformer-based architecture of the USE encoder as it was proven to yield better results. USE was implemented through its TF hub module [6]. For all pairs, each input question was embedded separately and then their combined embedding vector is formed as $(u, v, \mid u - v \mid, u * v)$, which is a concatenation of the premise and hypothesis vectors and their respective absolute difference and hadamard product. We experiment with both logistic regression and multilayer perceptron on top of the generated input representations. The MLP consists of a single hidden layer of 50 neurons using the adam optimizer and a batch size of 64.

[5] `https://scikit-learn.org/`

[6] The TF version of the USE model is available at `https://tfhub.dev/google/universal-sentence-encoder-large/3`

Run	**Model**	**Accuracy**	
		Dev	Test
1	Drop-out BERT AVG: 12 models with averaging ensemble	0.836	0.820
2	Stacking BERT 1: KNN	0.846	0.840
3	Stacking BERT 2: KNN followed by LR	0.847	0.847
4	Stacking BERT 2: (KNN/SVM/NB) followed by LR	0.849	0.852
5	Stacking BERT 2: Linear SVM followed by LR	0.846	0.823

Table 7.2: Results of our team runs on the MEDIQA challenge for the NLI task.

Run	**Model**	**Accuracy**	
		Dev	Test
1	USE embeddings with LR Classifier	0.770	0.584
2	USE embeddings with MLP Classifier (1 hidden layer with 50)	0.778	0.580

Table 7.3: Results of our team runs on the MEDIQA challenge for the RQE task.

7.5 Results & Discussion

7.5.1 Task 1: Natural Language Inference(NLI)

The best performing single BERT model achieved 0.828 for the evaluation set. Table 7.2 shows results of each model ensemble used for the NLI task. For the first run, we only averaged predictions generated by the ClinicalBERT model. The drop-out ensembling resulted in 12 models in total. For the second run, we used KNN classification over predictions from all trained BERT models. The remaining 3 runs use a variation classification models for the first level with a second level logistic regression classifier. We can observe consistent improvement from successive ensembling from one to two stacking levels. Our five runs showed substantial improvement in the performance over the original baseline with accuracy gain ranging from 10.6% to 13.8%. By the end of the challenge, 42 teams submitted a total of 143 runs to the NLI task. our top

performing submission ranked the 12th over all teams [7]. Its corresponding model could be viewed as a three-stage architecture with 2 level stacking ensemble as illustrated in figure 7.1.
All runs submitted relied solely on BERT text representations without any external features. Initially, we assumed that training our models with more than just embedding features should help classification and improve overall performance. We used the predictions generated by the drop-out averaging ensemble as extra features to further fine-tune a second-level BERT model. The model hyperparameters settings were the same as the best performing single base model. We did not find this experiment to yield any gains in the evaluation phase, compared to ensemble models, with only 0.815 accuracy for the development set. This was also affirmed post submission, with the release of the gold-labels. The accuracy for the test set was only 0.812.

7.5.2 Task 2: Recognizing Question Entailment (RQE)

Table 7.3 shows our two submitted runs for task 2. Even though our approach for this task was much simpler than task 1, we still managed to achieve a considerably good accuracy outperforming the baseline by 4.3%. The final results show that our team ranked the 23rd among all 54 participants[8]. Due to time constraints we were unable to fully investigate all models described in section 7.3, nor conduct a suitably thorough hyperparameter search for the MLP. However, we were able to conduct more evaluations post submission. We trained the inferSent Bi-LSTM encoder on the MedNLI data using GloVe embeddings. We then used the trained model to generate embeddings for the RQE data, and used the same MLP architecture to generate predictions. Despite the similarity of both tasks and the potential benefit from transfer learning, the model achieved an accuracy of 0.623 and 0.532 for dev and test sets respectively.

7.6 Conclusion

In this paper, we presented our solution for textual entailment detection in the clinical domain. Our proposed approach for the NLI task relies on BERT contextual embeddings features and machine learning algorithms such as KNN,

[7] Leaderboard for the NLI task: `https://www.aicrowd.com/challenges/mediqa-2019-natural-language-inference-nli/leaderboards` (accessed 1st of June 2019)

[8] Leaderboard for the RQE task: `https://www.aicrowd.com/challenges/mediqa-2019-recognizing-question-entailment-rqe/leaderboards` (accessed 1st of June 2019)

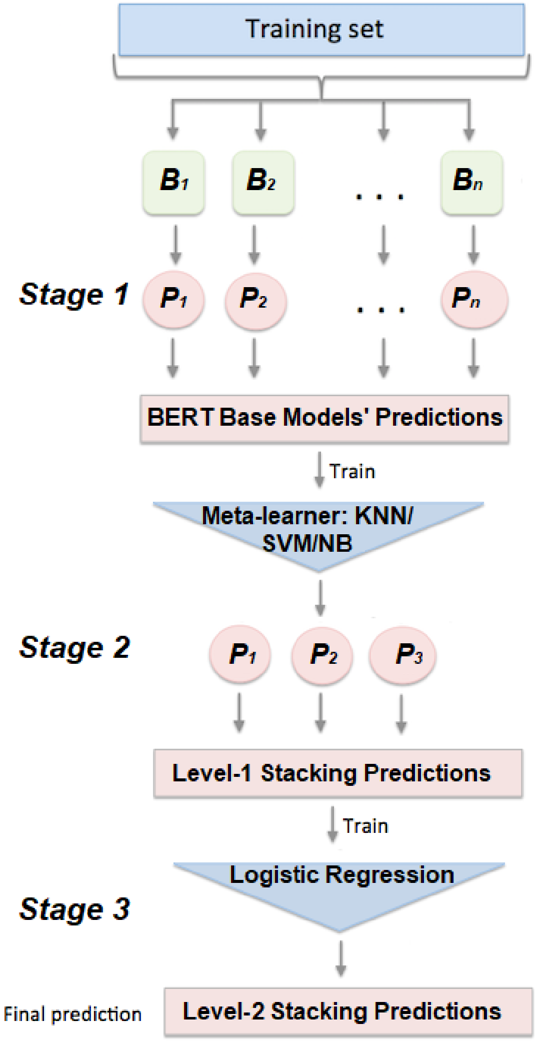


Figure 7.1: Overview of the ensemble architecture of the best run for the NLI task.

SVM and LR for ensembling. We use two different pre-trained BERT weights to train the base models and generate corresponding probabilities for the test set. Then, we adopt a 5-fold stacking strategy to learn and combine predictions. In the third and final level of the ensemble, we use a logistic regression over the outputs from level-1 stacking, to predict the final class labels. A future extension of our model is to use BERT in feature extraction mode instead of fine-tuning the end-to-end model on the MedNLI dataset. This would allow the selection of layers from which to extract embeddings and/or the combination of multiple layers. In the former scenario, different neural networks could be used to generate the base model predictions before applying ensemble techniques.

For the RQE task, we train an MLP classifier on top of USE embeddings. The results obtained were promising, given the simplicity of the model. More complex and deeper networks could be employed with the combination of USE embeddings. We also experimented with transfer learning by training the inferSent model on MedNLI before fine-tuning on the RQE corpus. While this approach did not improve the results, we aim at further investigating other inferSent architectures and training on clinical word embedding.

8 Computer-assisted Relevance Assessment of Literature

It is becoming more challenging for health professionals to keep up to date with current research. To save time, many experts perform evidence syntheses on systematic reviews instead of primary studies. Subsequently, there is a need to update reviews to include new evidence, which requires a significant amount of effort and delays the update process. These efforts can be significantly reduced by applying computer-assisted techniques to identify relevant studies. In this study, we followed a "human-in-the-loop" approach by engaging medical experts through a controlled user experiment to update systematic reviews. The primary outcome of interest was to compare the performance levels achieved when judging full abstracts versus single sentences accompanied by Natural Language Inference labels. The experiment included post-task questionnaires to collect participants' feedback on the usability of the computer-assisted suggestions. The findings lead us to the conclusion that employing sentence-level, for relevance assessment, achieves higher recall.

This work was originally published as:
Tawfik, N. S., & Spruit, M. R. (2020a) Computer-assisted Relevance Assessment: a Case-study of Updating Systematic Reviews. *Applied Sciences*, *10*(8), 2845.

8.1 Introduction

Relevance is a fundamental concept in Information Retrieval (IR) (Mizzaro, 1997). The entire search process revolves around the relevance concept where the effectiveness of a given system is measured by its ability to satisfy the user information needs. Defining relevance is, on its own, an active research area that started around 1959 when Brian Vickery discussed the distinction between relevance to a subject or a topic and the user relevance (Saracevic, 2007; Vickery, 1959a, 1959b). There are two main aspects of the IR process: system-driven and user-based aspects. The former focuses on finding information resources that match the user's search query and ranking them, while the latter targets the user's decision on assessing the relevance of the retrieved document. The main difference between both approaches is that, in the first one, the goal is to extract relevant data from a bigger pool of data, while the second's goal is to determine the usefulness of the extracted data. This usefulness is usually subjective as it depends on the user's information needs and varies across people (Janes, 1994).

A widely adopted measure of effectiveness in the IR domain is the construction of benchmark test collections on specific topics; a large document set with each document manually labeled as relevant or irrelevant. An automated process computes the similarity between IR system output and the test collection labels. This provides flexibility as system-driven evaluation can then be repeated and extended (Sanderson, 2010). On the other hand, conducting user studies that involve human participants who test IR systems is an important method. This method has the advantage of collecting and analyzing important factors such as user judgments, user behavior, user satisfaction, and perceived problems. Despite the accompanied issues such as the difficulty in subjects recruitment, cost and, reproducibility (Kelly & Kelly, 2009), in this study, we focused on evaluating the relevance assessment through the latter, user-based approach.

In the current era of big data, assessing relevance while maintaining a high recall is a crucial problem for information retrieval systems. In many applications, thousands of human assessments are required to achieve it, which is costly both at the time and the effort level. Examples of such applications include electronic discovery (e-discovery) especially in the legal domain, evidence-based collection in the medical domain for systematic reviews, and even construction of test collections for information retrieval tasks. High-recall driven systems are designed to reduce the cost of reliance on human input in assessing relevant documents while maintaining reasonable assessment effort.

With the exponential increase of published medical literature and the wide adoption of electronic medical records, there is an increasing need for information retrieval systems tailored to the needs of experts in searching medical textual data. Manual relevance assessment in medical IR is found to be a time demanding and cognitively expensive task (Koopman & Zuccon, 2014). Another challenge in medical IR is the variations in judgment agreements among assessors according to their expertise level, and their understanding of the document. This could profoundly impact document relevance assessment and overall retrieval performance (Tamine & Chouquet, 2017). While relevance assessment is subjective, some factors also influence the judgment accuracy, such as the search topic, the language level, the documents length, and the review time and speed.

This research aims at bridging the gap between the text mining community and the medical community. It investigates how to speed up the task of information retrieval in general, and when high levels of a recall are required specifically. We focus on one of the daily challenges in clinical practice; the relevance assessment of biomedical literature. We designed and conducted an experiment to analyze the time, effort, and decisions of medical experts in judging the relevance of scientific articles. To the best of our knowledge, there is no existing controlled user study, in the biomedical domain, that explores the relevance judgment behavior of expert assessors while varying the content-length of the previewed documents. We focused on the performance achieved by only showing the conclusion sentences paired up with computer-assisted information as opposed to showing the full abstract. The objective was to determine if acceptable recall could be achieved in a shorter amount of time by viewing less and taking advantage of extra information using language inference models. More specifically, we investigated the speed-up of the relevance assessment process, by showing less, while maintaining the quality of the judging.

8.2 Related Work

The research on relevance assessment involves many aspects, however, two critical factors are important for optimizing the assessment process: the time factor and the content-length factor. The time factor consists of measuring the amount of time that assessors need to perform relevance judgments, while the content-length factor denotes the percentage of text shown to users ranging from a single sentence to the full-text document.

Maddalena et al. highlighted the effect of time on the quality of the judging

(Maddalena et al., 2016). The authors reported that applying time constraints could reduce the costs associated with crowdsourcing with no loss of quality. Their findings show that only 25–30 s are enough to achieve top judgment quality for a single document. Wang and Soergel conducted a comparative user study in the legal e-discovery domain between participants with and without law background (J. Wang & Soergel, 2010). They investigated different parameters in relevance assessment, including speed, accuracy, and agreements. They found no significance between both groups in the speed or the accuracy when participants are provided with correct guidelines to judge document. In another study that included eight students and used the TREC e-discovery test collection, the authors conducted a correlation analysis and found that speed highly correlates to the perceived difficulty of the document (J. Wang, 2011). However, no correlation was observed between the judgment accuracy and speed (with the exception of one topic). There was a notable variation among assessors in their judgment speed ranging from 13 to 83 documents per hour with an average of 29 documents. It was also found that assessors judged non-relevant documents about as fast as relevant documents.

On the other hand, the effect of content-length on the relevance judgment accuracy was also investigated through user-based experiments. In (Tombros, Sanderson, & Gray, 1998), Tombros and Sanderson aimed at minimizing the need to refer to full-text documents by providing documents' summaries to support the retrieval decisions. Their experiment included 20 participants, and summaries were approximately 15% of the original text length mounting up to almost five sentences. The authors also introduced a 5-min time limit for each participant to identify the relevant documents. Their experimental results show that the user group that relied on summaries for judgment identified more relevant documents more accurately than the group who had access to the full document. In (Sanderson, 1998), the author investigated user-directed summaries, i.e., summaries biased towards the user's needs within the context of Information Retrieval. The paper highlights, amongst other conclusions, that users judge documents relevancy from their corresponding summaries almost as accurately as judging from their full text. It also reveals that assessing the full document needed, on average, 61 s while the document summary could be judged in only 24 s. Similarly, a study on the relevance judgment of news articles used user-biased snippets of documents in a controlled experiment. Their approach restricted shown snippets to a maximum of 50 words, an equivalent of two sentences or fewer. In that study, Smucker and Jethani found that the average time to judge a document was 15.5 and 49 s for summaries and full-documents, respectively. They also reported that the probability of judging summaries relevant is lower than documents (Smucker

& Jethani, 2010). Zhang et al. recruited 50 users to evaluate the use a single paragraph as relevance feedback in a Continuous Active Learning (CAL) settings (Zhang et al., 2018). Users were allowed only 1 h to find as many relevant documents as possible using an in-house built IR system. As expected, more relevant documents were retrieved by viewing document excerpts as opposed to full documents. In another attempt to reduce annotation overhead, the same authors suggested using single sentences for feedback in CAL (Zhang, Cormack, Grossman, & Smucker, 2019). The single sentence could contain sufficient information for an assessor to provide relevance judgment with an acceptable level of performance. Their results were based on a simulation study and not a human experiment, where an IR system that mimics different aspects of human interactions generates the relevance feedback.

More recently, Rahbariasl and Smucker (Rahbariasl & Smucker, 2019) conducted a user experiment that combined different time limits with summaries and full documents. The study design included seven topics from the 2017 TREC Common Core track with 60 enrolled participants. They were given 15, 30, and 60 s of time limits to judge document relevancy with respect to a topic. The results show that, as the time limit increases, the average time to judge a document increases regardless of whether it is a full document or a summary. Overall, neither the time limits nor the document type had a statistical significant effect on accuracy. The authors suggested that employing summaries for speeding relevance judging is a better solution than imposing time limits. Their suggestion was based on the post-experiment feedback questionnaire, which shows that, with the maximum time limit of 60 s, participants enjoyed the experience with no stress involved. In that setting, they were able to judge a document summary in 13.4 s and a full-text document in 22.6 s while achieving a performance of 0.73 and 0.74 for summaries and full-text, respectively.

In the medical domain, there is scattered work on the influence of time and length variables on the assessment task based on expert judges. Existing work on relevance assessment is mostly related to TREC and CLEF evaluation challenges through the means of constructing test collections and reporting annotators' experience (Kanoulas, Li, Azzopardi, & Spijker, 2017; Kanoulas, Li, Azzopardi, Spijker, & others, 2018; Roberts et al., 2017). In a previous work by Koopman and Zuccon (Koopman & Zuccon, 2014), the authors contradicted the findings of common intuition and previously mentioned studies as they reported that time spent to assess relevance is not related to document length but is more query-dependent. Another related study aims at investigating and improving how to effectively use search systems to extract scientific medical literature (SML) of interest. The study evaluates the impact of imposing time constraints on clinicians in answering medical questions with the help of an SML search system. The study is still ongoing, and results from participants testing is not published yet (Van Der Vegt, Zuccon, Koopman, & Deacon, 2019).

8.3 Materials and Methods

We assessed the effectiveness of the suggested hypothesis by means of user studies and domain test data. Generally, a test collection consists of either real or artificial data, generated by the examiner(s). We performed experimentation on a real-world case study that emulates the update process of biomedical systematic reviews (SRs) according to newly published literature.

8.3.1 Case Study: Updating Systematic Reviews

Scientific literature remains the primary source of information for clinicians and experts in the medical domain. The need for information synthesis is becoming indispensable, specifically with the adoption of the precision medicine concept into practice. Not only do health professionals need to keep up to date with current research, but they have to curate relevant information linking genomic variants to phenotypic data to make informed, personalized clinical decisions per patient. This makes finding relevant information more challenging, even with the availability of user-friendly search engines such as PubMed and MEDLINE that facilitate access to available publications. Additionally, the rate of publication of biomedical and health sciences literature is increasing exponentially. As an indication, it has been estimated that the number of clinical trials increased from 10 per day in 1975 to 55 in 1995 and 95 in 2015. In 2017, the PubMed repository contained around 27 million articles, 2 million medical reviews, 500,000 clinical trials, and 70,000 systematic reviews (Catillon, 2017a). Oftentimes, medical experts do not seek answers to their questions due to time restrictions, or they suspect that no useful outcome will result from their search (Ely, Osheroff, Chambliss, Ebell, & Rosenbaum, 2005).

This tsunami of data has led time-pressured clinicians to perform evidence syntheses on systematic reviews instead of primary studies. Systematic reviews are comparative effectiveness research studies that use explicit methods to find, evaluate, and synthesize the research evidence for a research question (Morton, Berg, Levit, Eden, & others, 2011). Extracting information from reviews better suits experts' needs as there are fewer hits to curate while comparing findings across different studies efficiently (Pieper, Antoine, Neugebauer, & Eikermann, 2014).

Subsequently, there is a need to regularly check reviews to keep them up-to-date with current studies (Bashir, Surian, & Dunn, 2018; Pieper et al., 2014). Shojania et al. (Shojania et al., 2007) conducted a survival analysis on a sample of 100 systematic quantitative reviews to estimate the average time for changes in evidence. Their findings show that newly published evidence was available for 23% of the sample within two years and for 15% within one year. Moreover, 7% of the reviews were already out-of-date by the time of publication. Creating or updating systematic reviews is a time-consuming task mostly because it involves creating queries, fetching results, manual screening of candidate studies, assessing articles' relevance to the topic, and satisfying the inclusion criteria. The process of creating or updating SRs typically requires 6-12 months of effort, with the main expense being personnel time (A. M. Cohen, Ambert, & McDonagh, 2010). Many technology-assisted models have been proposed to reduce the efforts of the task. The models mainly rely on information extraction and text mining techniques in an automated or semi-automated approach.

We designed a user-experiment to analyze the relevance assessment task within the process of updating systematic reviews. In that task, relevance assessment refers to the task of determining the relevance of an article with respect to a review topic. It can be interpreted as a measurement in which assessors judge the pertinence of a document to a chosen topic. The primary outcome of interest was the performance levels achieved when replacing the full abstract with only a single sentence enriched with textual inference labels.

8.3.2 Medical Natural Language Inference

This research partly builds on our previous work, in which we investigated the Natural Language Inference (NLI) task in the biomedical domain. Given two snippets of text, *Premise* and *Hypothesis*, textual inference determines if the meaning of H can be inferred from that of P (Dagan, Roth, Sammons, & Zanzotto, 2013b). Our work uses scientific literature to detect biological conflicts and literature inconsistencies between reported medical findings. The proposed NLI model follows a siamese Deep Neural Network (DNN) architecture with three layers. The input text is embedded using InferSent (Conneau et al., 2017), a sentence encoder trained on NLI sentence pairs, additional semantic features such as sentence length, cosine similarity, and contradiction-based features such as negation, modality, polarity, and antonyms. The features are integrated into the third layer of the network to improve performance. For all input pairs, the model accordingly assigns an inference label to each pair: *Entailment, Neutral, Contradictory.* The training data consist of 2135 claim pairs extracted from abstracts, on the topic of cardiovascular diseases. More details of Biomedical NLI models can be found at (N. Tawfik & Spruit, 2019; N. S. Tawfik & Spruit, 2019b).

8.3.3 Experiment Design

The experiment session comprised two main tasks carried out in sequence, namely *Abstract-View* task and *Sentence-View* task. It also included pre and post task questionnaires for demographics information and feedback survey. The experiment was created and hosted on the online research platform Gorilla (`https://gorilla.sc/`). Gorilla is a commercial web platform originally designed for behavioral science tests through questionnaires and cognitive tests. It offers GUI construction and visual programming, as well as allows setting time limits and collection of reaction times data (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2019). With these features, Gorilla was more suitable to deploy in our experiment as opposed to regular surveying tools with limited functionality when analyzing human behavior.

Data Collection

As mentioned, the main focus of this study was to measure how human performance changes according to the length of evidence shown. To carry such an experiment, we had to collect a set of documents with actual relevance pre-judged by domain experts. In both tasks of the experiment, users viewed a list of clinical studies associated with a systematic review extracted from the Cochrane Database of Systematic Reviews (CSDR) (`https://www.cochranelibrary.com/cdsr/about-cdsr`). The CSDR is a leading database for systematic reviews helping medical experts in the decision-making through evidence-based information. A Cochrane review identifies, appraises, and synthesizes evidence related to a specific research question given pre-specified eligibility criteria. Generally, to collect studies for systematic reviews, experts construct complex boolean queries that include logical operators such as AND, OR, and NOT to maximize the recall.

The results from the query search pass through two elimination phases, abstract-screening and full-text screening. The abstract-screening involves a review of article titles and abstracts to give initial feedback on the relevancy of the study. This is followed by a full-text screening that consists of a thorough analysis to decide whether the study meets the inclusion criteria or not.

Each systematic review included in the Cochrane database has its bibliography divided into three: *Included, Excluded, and Additional* sections, with references to other articles. Studies in the *Included* section are the ones found relevant to the systematic review based on the full content of the study, in other words, after passing through both elimination stages, whereas the *Excluded* section includes studies that were initially considered relevant in the abstract screening stage but later dismissed in the full-text screening stage. On the other hand, the *Additional* section includes references that are neutral and considered irrelevant to the study.

To fully assess the human judgment, the list of documents to be reviewed by the participants needs to be uniform, i.e., to maintain a good balance between relevant and irrelevant documents included in the list. Another constraint to account for while choosing the candidate documents is the difficulty level in assessing the documents; it should be as tricky as it is for experts when updating a systematic review in reality. With these goals in mind, we selected six published intervention systematic reviews, from the heart and circulation category, that are available in the CSDR. The chosen reviews are assigned the *New Search* tag, which indicates that the authors of the review published an updated version.

We limited the scope of the experiment to the abstract screening stage, and hence references in the *Included* and *Excluded* sections are both considered relevant studies at the abstract level. Similarly, all references to studies in the *Additional* section are deemed irrelevant. We note that we excluded any reference to studies not related to the biomedical field, such as studies with general guidelines on how to conduct systematic reviews or interpret bias. In some cases, we re-submitted the boolean query, if provided, to the corresponding medical database and the retrieved studies were also considered irrelevant even if not included in the *Additional* section. Our list of documents remains, however, random and un-ordered. Assessing and understanding human judgments and responses to ranked lists of documents is more complicated, and we leave it for future work. For each systematic review included in the dataset, we collected the following information: (1) review title; (2) main finding of the original systematic review; (3) PubMed ids (PMIDs) of relevant studies (references from the *included* and *excluded* sections of the updated version); and (4) PubMed ids (PMIDs) of irrelevant studies (references from the *Additional* section of the updated version or from re-submitting the query). The average percentage of relevant studies in the collected set is 45.4% of the total number of PMIDs included. Table 8.1 shows the distribution of the relevant and irrelevant documents of all the topics in the set.

To prepare the data for the *Abstract-View* task, PMIDs of relevant and irrelevant references were further used to download the full abstract of the studies through the Biopython library and the NCBI/PubMed API. For the *Sentence-View* task, we took advantage of the fact that abstracts of the studies follow a structured format with sections such as Background, Objectives, Results, and Conclusion. We extracted the first sentence of the Conclusion section and coupled it with the finding of the original review. The sentence pair served as input to the in-house NLI model described in Section 8.3.2 to generate an inference label. Additionally, each article has a relevancy value (True for relevant studies and False for not relevant) according to which reference section they belong to.

Table 8.1: Distribution of *Relevant* and *Irrelevant* articles in the data collection.

Review ID	**Updated Review ID**	**Relevant**	**Irrelevant**	**Total**
21249668	29240976	11	12	23
26308931	29667726	11	14	25
15846608	23235577	12	14	26
15266480	22293766	10	13	23
22696339	26123045	10	15	25
21901719	26934541	11	10	21

Participants

We aimed to recruit medical experts with knowledge in conducting systematic reviews. They were recruited by word of mouth and personal contact through emails. Aside from access to a computer or laptop with a stable Internet connection, there were no specific exclusion criteria for the study. Participants received a detailed email with the rules and information that needed to introduce the study, along with the study URL. In the invitation email, participants were asked to complete the whole experiment in a single sitting. They were also encouraged to ask for help, via email, if they had any queries or problems. All users were requested to complete a consent form prior to taking part in the experiment. Each participant completed the experiment at their convenience within two weeks from the invitation mail.

Interface

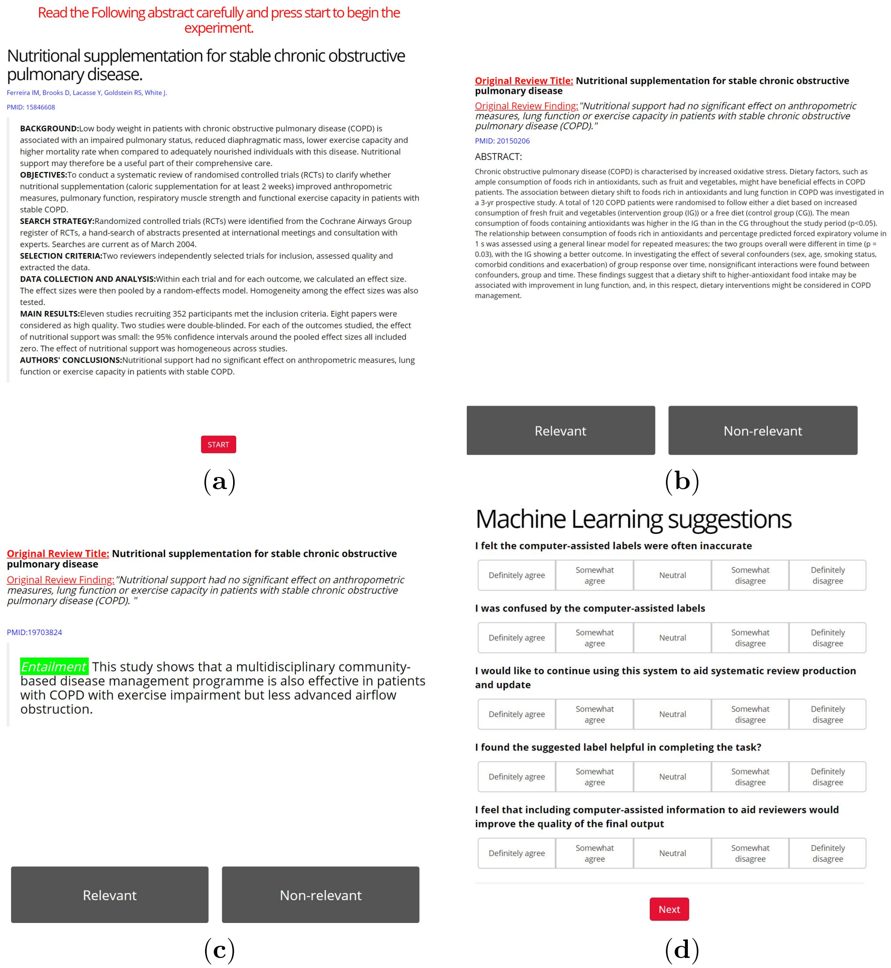
We plotted the experiment interface to become as much intuitive and visually pleasant as possible, and to engage participants in the document judgment task. It shows them one document at time in either tasks. For the main tasks, the screen shows either the full abstract or the conclusion sentence of the candidate study to be judged along with its corresponding PMID. We followed the standard way to collect relevance judgments where participants only could make binary judgments by clicking on relevant or irrelevant buttons. Through the whole study, participants could see the the title and original finding of the main review so that they do not forget the main topic. Figure 8.1 shows an example of the user interface for the main tasks.

Procedures and Task Description

The designed web-based experiment for relevance assessment was divided into four main parts. Figure 8.2 illustrates the experiment phases.

Consent and Demographic data: At the beginning, all participants were requested to complete a consent form prior to taking part in the experiment. Next, they were asked to fill in an eleven-question survey to capture demographic data, medical background, and expertise.

Main Experiment: Following the demographic questionnaire, the main experiment started, which consisted of two tasks: the *Abstract-View* task and the *Sentence-View* task. Once the first task started, the participant reviewed specific instructions on how to complete the task. They were reminded that the experiment involves a time limit and that they should carefully read the original systematic review before moving to the assessment stage and starting the countdown. The shown instructions motivated the participants to optimize their speed and recall by instructing them to *"Try to find as many relevant documents as possible in the 15 min while still making as few mistakes in judging the documents' relevance as possible"*. Such language was also employed in similar user experiments in other domains (Smith & Kantor, 2008; Smucker & Jethani, 2010). First, the participant read the abstract of the original systematic review. In this step, participants acquired knowledge about the topic and also gained insights on the inclusion criteria of the review. This step was not included in the time limit since relevance judgments were based on the participants' understanding of the topic. Next, a series of abstract texts was presented to the participant, one at a time for a duration of 15 min. To complete the task, the user selected a single judgment (Relevant,



Read the Following abstract carefully and press start to begin the experiment.

Nutritional supplementation for stable chronic obstructive pulmonary disease.

Ferreira IM, Brooks D, Lacasse Y, Goldstein RS, White J.

PMID: 15846608

BACKGROUND: Low body weight in patients with chronic obstructive pulmonary disease (COPD) is associated with an impaired pulmonary status, reduced diaphragmatic mass, lower exercise capacity and higher mortality rate when compared to adequately nourished individuals with this disease. Nutritional support may therefore be a useful part of their comprehensive care.
OBJECTIVES: To conduct a systematic review of randomised controlled trials (RCTs) to clarify whether nutritional supplementation (caloric supplementation for at least 2 weeks) improved anthropometric measures, pulmonary function, respiratory muscle strength and functional exercise capacity in patients with stable COPD.
SEARCH STRATEGY: Randomized controlled trials (RCTs) were identified from the Cochrane Airways Group register of RCTs, a hand-search of abstracts presented at international meetings and consultation with experts. Searches are current as of March 2004.
SELECTION CRITERIA: Two reviewers independently selected trials for inclusion, assessed quality and extracted the data.
DATA COLLECTION AND ANALYSIS: Within each trial and for each outcome, we calculated an effect size. The effect sizes were then pooled by a random-effects model. Homogeneity among the effect sizes was also tested.
MAIN RESULTS: Eleven studies recruiting 352 participants met the inclusion criteria. Eight papers were considered as high quality. Two studies were double-blinded. For each of the outcomes studied, the effect of nutritional support was small: the 95% confidence intervals around the pooled effect sizes all included zero. The effect of nutritional support was homogeneous across studies.
AUTHORS' CONCLUSIONS: Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable COPD.

(**a**)

Original Review Title: Nutritional supplementation for stable chronic obstructive pulmonary disease

Original Review Finding: *"Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable chronic obstructive pulmonary disease (COPD)."*

PMID: 20150206

ABSTRACT:

Chronic obstructive pulmonary disease (COPD) is characterised by increased oxidative stress. Dietary factors, such as ample consumption of foods rich in antioxidants, such as fruit and vegetables, might have beneficial effects in COPD patients. The association between dietary shift to foods rich in antioxidants and lung function in COPD was investigated in a 3-yr prospective study. A total of 120 COPD patients were randomised to follow either a diet based on increased consumption of fresh fruit and vegetables (intervention group (IG)) or a free diet (control group (CG)). The mean consumption of foods containing antioxidants was higher in the IG than in the CG throughout the study period (p<0.05). The relationship between consumption of foods rich in antioxidants and percentage predicted forced expiratory volume in 1 s was assessed using a general linear model for repeated measures; the two groups overall were different in time (p = 0.03), with the IG showing a better outcome. In investigating the effect of several confounders (sex, age, smoking status, comorbid conditions and exacerbation) of group response over time, nonsignificant interactions were found between confounders, group and time. These findings suggest that a dietary shift to higher-antioxidant food intake may be associated with improvement in lung function, and, in this respect, dietary interventions might be considered in COPD management.

(**b**)

Original Review Title: Nutritional supplementation for stable chronic obstructive pulmonary disease

Original Review Finding: *"Nutritional support had no significant effect on anthropometric measures, lung function or exercise capacity in patients with stable chronic obstructive pulmonary disease (COPD). "*

PMID:19703824

Entailment This study shows that a multidisciplinary community-based disease management programme is also effective in patients with COPD with exercise impairment but less advanced airflow obstruction.

(**c**)

Machine Learning suggestions

I felt the computer-assisted labels were often inaccurate

Definitely agree | Somewhat agree | Neutral | Somewhat disagree | Definitely disagree

I was confused by the computer-assisted labels

Definitely agree | Somewhat agree | Neutral | Somewhat disagree | Definitely disagree

I would like to continue using this system to aid systematic review production and update

Definitely agree | Somewhat agree | Neutral | Somewhat disagree | Definitely disagree

I found the suggested label helpful in completing the task?

Definitely agree | Somewhat agree | Neutral | Somewhat disagree | Definitely disagree

I feel that including computer-assisted information to aid reviewers would improve the quality of the final output

Definitely agree | Somewhat agree | Neutral | Somewhat disagree | Definitely disagree

(**d**)

Figure 8.1: Screenshots of the user interface for both tasks and questionnaire in the experiment: (**a**) Original systematic review; (**b**) *Abstract-View* task; (**c**) *Sentence-View* task; and (**d**) Feedback questionnaire.

or Non-relevant). Finally, the system asked the user if this systematic review qualifies for a *Conclusion Change*, which indicates that modifications in the original findings and/or a more precise conclusion should be issued as per the Cochrane Library guidelines. The experts responded based on articles they judged as relevant and their corresponding conclusions with respect to the original SR. At the end of the first task, the system moved the participant to the second task of the experiment. The same instructions were once again shown to participants with modifications to the sentence-level task. Once again, the participant read a systematic review abstract and completed the same task in judging related documents as relevant and irrelevant. However, at that stage, the participant only viewed a single sentence and a label that indicated its inference relation with the original finding.

Each task corresponded to one of the six collected systematic reviews. We employed the built-in randomization procedure of Gorilla for assigning systematic reviews to participants. We enforced a balanced design so that all six reviews were equally included in the review and set a constraint to show different systematic review for each user per task, i.e., no duplication of documents between *Abstract-View* and *Sentence-View* tasks. At time-out, the screen was blocked, and the participant was taken to the task completion screen to enter their final details. We collected all the users' relevance judgments and time spent to judge each document throughout both tasks. All participants acknowledged the time allocation at the start of each task but no countdown timer was visible during the task. This provided a balance between making the participant aware of the time allocated for each task without distraction that might divert their attention from from the task.

Feedback: After completing both tasks, participants answered an exit questionnaire to collect their feedback and overall experience. To evaluate usability, we collected qualitative feedback on the experiment and the subjective perception of system usability. The feedback questionnaire follows the IBM Computer System Usability Questionnaire (CSUQ) (Lewis, 1995); however, the questions were adapted to fit the scope of our experiment. Additionally, we added specific questions focused on the computer-assisted suggestions to investigate the usability of the NLI independently. Finally, we asked participants to optionally provide extra feedback on the experiment via a free-text form.

Participants were able to withdraw from the study at any time by closing their browser. The whole experiment required participants to work for almost 1 h; each task required 15 min in addition to 10 min for reading the original review. We estimated that responding to the demographics and feedback survey would require 10 min. The Gorilla platform enables setting a time

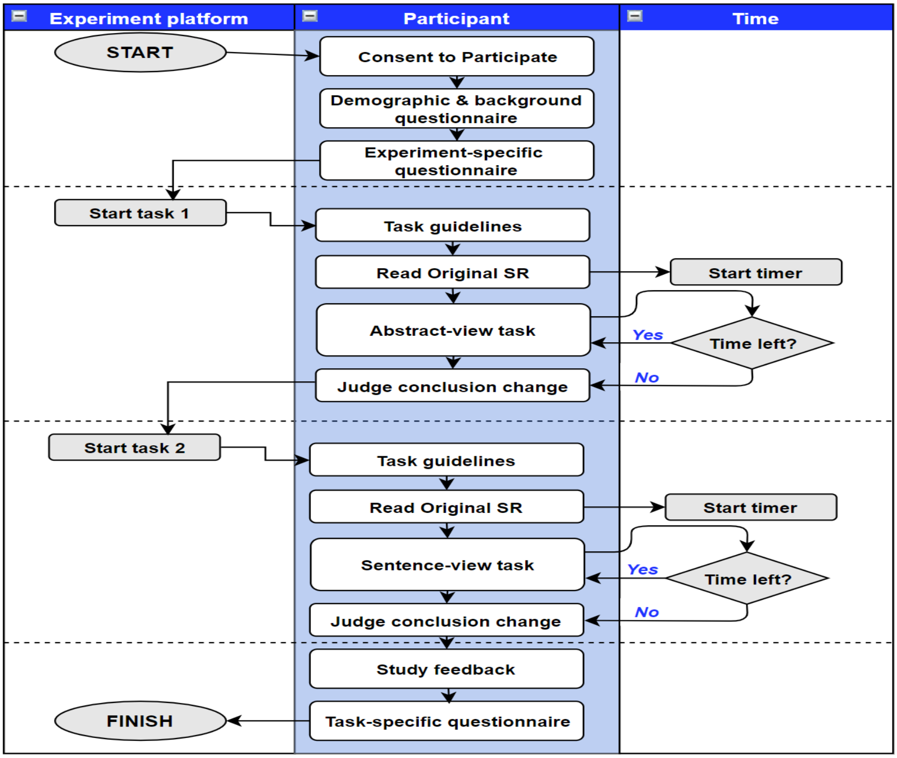


Figure 8.2: Process flow diagram for both tasks of the experiment.

limit per participant to reach the final checkpoint. For convenience, we set the time limit to 90 min to allow for breaks between tasks. Participants who exceeded the time limit were eliminated from the study.

8.4 Results

8.4.1 Recruitment

Twenty-two researchers (4 male and 18 female) voluntarily participated in the study. Their age mostly fell in the 46–65 years category with 12 participants, followed by 9 participants in the 25–45 years category and only 1 participant below 25 years. All participants had masters or doctorate degrees in different biomedical fields. Most of them were affiliated to Egyptian universities, either as students (5 out of 22 participants) or as academic staff, i.e., lecturers, research assistants, or professors (17 out of 22 participants). The pre-experiment questionnaire aimed at collecting information about the participants' familiarity with biomedical literature curation and their perception of the search process. The majority of the participants use PubMed and MEDLINE (10 and 9 out of the 22 participants, respectively); alternatively, some use Google scholar for their searches (3 out of 22 participants). Table 8.2 summarizes the participants demographics. Almost all participants tend not to use advanced features of search engines or other computer-assisted evidence synthesis tools (95% of the participants). However, they showed interest in learning on how computer-assisted system could speed up their search. Eight participants were involved in preparing and publishing systematic reviews. As far as their perception about the different tasks for searching and validating biomedical evidence, 15 of the participants think that assessing relevance is the most challenging task, while 7 believe that the task of building queries is more complicated.

Table 8.2: Results of the relevance assessment experiment.

Measure		**Sentence-View**
Gender	Female	18
	Male	4
Age Group	18–24	1
	25–45	9
	46–65	12
Education	Bachelor	1
	Master	4
	Doctoral	17
Frequency of using medical IR tool	Weekly	15
	Monthly	6
	Yearly	1
Medical IR tool	PubMed	9
	MEDLINE	10
	Google scholar	3

8.4.2 Performance

Relevance is subjective even at the expert level, different judgments could be inferred for the same article. Accordingly, it is essential to define, within the context of the experiment, what documents count as relevant, and why. In the case study of systematic reviews update, relevance judgments are determined by the authors of the review through the reference sections, as described in Section 8.3.3. In our experiment, we considered the authors' judgments as the gold-standard annotations since they initially set the inclusion and exclusion criteria for each review. For each article, we counted documents as relevant if both the participant and the author of the systematic review agree on their relevancy, i.e. the participant clicked on the "Relevant" button, and the article is in the *included* or *excluded* reference sections, and similarly for the irrelevant documents. The correctly judged relevant and irrelevant sets are denoted S_R and S_{IRR}, respectively.

To compare the difference in relevancy judgment between sentence and abstract views, a logical measure to use is the number of correct relevant articles reported by the user, S_R, for both tasks. We interpret accuracy based on correctly assessed documents, namely the true positives (TP) and true negatives (TN), which indicate if the participants agree or disagree with the ground truth judgments. S_R represents the TP as the participant judgment aligns with the author judgment positively while S_{IRR} represents TN.

$$Accuracy = \frac{S_R + S_{IRR}}{N} \tag{8.1}$$

where N is the total number of articles per systematic review. Similar to the legal e-discovery task, conducting a systematic review aims at finding all available relevant documents. Missing documents could lead to legal issues for e-discovery or could affect the conclusions reported by the systematic review. Therefore, there is a need for a better suited measure of performance for this case study, such as recall, to measure the fraction of all relevant documents found by participants. In our evaluation, we calculated recall by dividing the number of correctly judged relevant documents (S_R) by the total number of relevant documents (R) available for each systematic review.

$$Recall = \frac{S_R}{R} \tag{8.2}$$

In statistical analysis of binary classification, the F-score, or F-measure, is a measure of a test's accuracy that combines both recall and precision to compute the score. The higher is the score, the better, reaching its best value at 1. The general formula to calculate the F-measure requires a positive real beta parameter (β) as follows:

$$F_\beta = (1+\beta^2)\frac{Precision.Recall}{(\beta^2.Precision) + Recall} \tag{8.3}$$

Increasing the beta parameter consequently means emphasizing recall over precision. Given the recall-oriented nature of the task, we report the F2 score, which weights the recall twice as much as precision as opposed to the F1 score that portrays the harmonic mean of the precision and recall. Additionally, we tracked the time needed to judge each article in both the *Abstract-View* and *Sentence-View* tasks. As shown in Table 8.1, each topic has a different number of relevant documents. To minimize the skewness of reported relevant articles, we normalized each systematic review according to the number of relevant documents in its set. Table 8.3 shows the values for the computed measures; the corresponding values are reported on the average basis among all participants, across all systematic reviews from the data collection.

Table 8.3: Results of the relevance assessment experiment.

Measure	**Abstract-View**	**Sentence-View**
Seconds needed to judge each article	37.45	19.85
Accuracy of the relevance-assessment task	0.56	0.66
Recall of the relevance-assessment task	0.59	0.66
F2-score for the relevance-assessment task	0.57	0.64
Number of viewed documents	17.04	18.42
Number of correctly judged documents	11.00	12.33

Feedback

The participants found that the interface was user-friendly for the experiment and that the guidelines for each task were well explained, which made the judging tasks enjoyable. The results are based on the post-experiment questionnaire after both tasks. Almost half of the participants agreed that time limits imposed extra stress on the judgment process with an inter-participant average agreement of 2.67, while four thought that the time factor was neutral. In the second part of the survey, participants were requested to directly compare the *Abstract-View* and *Sentence-View* tasks through the following questions:

(Q1) How difficult was it to determine if a document was relevant or not to the systematic review?

(Q2) How would you rate your experience of judging the relevance of documents through the system?

(Q3) How confident did you feel while doing the task?

(Q4) How accurate do you think you have judged the documents?

We used a five-point Likert scale to map each question to values 1–5 with the most negative answer mapped to 1 and the most positive answer mapped to 5. The average response to the task-specific questions are shown in Figure 8.3. The results show that participants felt more confident and believe their answers were more accurate in the *Sentence-View* task. We also assessed the usability of the computer-assisted labels, based on the NLI model, in helping experts judging the relevance through six statements in the final part of the survey:

(S1) I felt the computer-assisted labels were often inaccurate

(S2) I was confused by the computer-assisted labels

(S3) I would like to continue using this system to aid systematic review production and update

(S4) I found the suggested label helpful in completing the task

(S5) I feel that including computer-assisted information to aid reviewers would improve the quality of the final output

(S6) I would use a similar system to to check updated information regarding a medical case

The user could express a level of agreement to the statements ranging: Definitely agree, Somewhat agree, Neutral, Somewhat disagree, and Definitely disagree. Table 8.4 shows the corresponding responses for each statement. The results show that participants favored the computer-assisted information and would use a similar system for curating evidence-based medicines. This is also supported by the responses to the final question in the survey where users were asked if they would recommend this system to a friend and 98% answered positively.

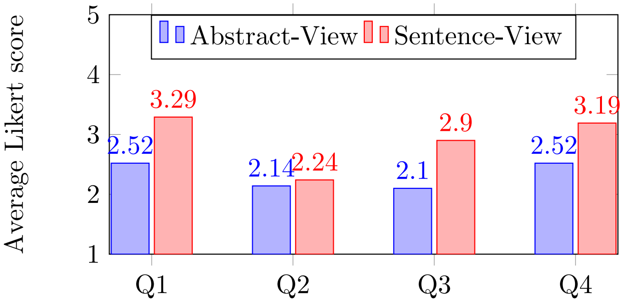


Figure 8.3: Task-specific questionnaire results. The participants used values 1–5 with the most negative answer mapped to 1 and the most positive answer mapped to 5.

Table 8.4: Participants agreement on the usability of the NLI labels.

	Definitely Agree	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Definitely Disagree**
S1	0	5	4	13	0
S2	3	8	3	6	1
S3	10	6	5	1	0
S4	5	14	2	1	0
S5	15	6	0	1	0
S6	14	6	2	0	0

8.5 Discussion

Similar to tombros1998advantages, we presume that the total number of experts who participated in the experiment (22 participants) is enough to draw significance to any results obtained. The following summarizes the insights from our experiment:

Less is More: There was a difference in the judging behavior among the *Abstract-View* and *Sentence-View* tasks. The assessors' quality of relevance judging when shown single sentences and labels is not only as good as when shown full abstracts but also in most the cases better in terms of accuracy and recall.

Versatile reading behavior: In the *Abstract-View* task, some participants took their time to judge the relevancy and read the abstracts carefully. Others finished the task quite soon by skimming the abstracts. This is reflected in the relatively small difference in the value of the total number of viewed articles between both tasks. We also observed that the 15 minutes limit provided plenty of time for participants to judge the related documents, even for the *Abstract-View* task, given the small size of the dataset (only 20–25 articles per systematic review).

The gains of the semi-automated approach: Machine learning suggestions can be used to support relevance assessment for updating systematic biomedical reviews via a web-interface. Participants appeared to engage easily with the system, rated the system as very highly usable, and would likely continue using a similar one for their daily search and curation activities. More generally, the "human-in-the-loop" experiments, such as this study, are important to demonstrate any utility afforded by text mining and machine learning.

Quality of computer-assisted suggestions: The quality of the biomedical textual inference labels is important and needs to be very accurate so that users can trust and base their judgments upon it. The feedback results show that some of the participants were confused by the highlighted inference labels. This might be because they had little knowledge on how labels were generated and were unfamiliar with the NLI model or its accuracy. A possible feature when designing an information retrieval system is to offer an opt-in or opt-out option for showing suggestions and also providing a confidence score of the assigned labels.

8.6 Conclusions

In this study, we conducted a controlled user experiment to investigate the assumption that assessing document excerpts can reduce assessment time and effort and still achieve high recall. We designed a case study based on the process of updating systematic medical reviews. The case study comprises two tasks that display either full abstracts or conclusion sentences with computer-assisted labels for expert assessors to judge. Throughout each task, participants were requested to find as many relevant documents as possible within 15 min. We computed the proportion of correct judgments (i.e., the user relevancy decision is in agreement with the systematic reviews' authors) that were returned by the participants and the accuracy and recall levels in both tasks. Participants additionally completed two brief questionnaires: one at the beginning of the study, which consisted of eleven questions concerning their level of experience and demographics, and one at the end to collect the participants' feedback on the experiment in general and on the usability of the computer-assisted labels specifically. Participants were academic professionals, with a high-level of expertise, familiar with the process of conducting a systematic review. The results of the controlled user study show that assessors were able to judge more relevant articles within the time limit in the *Sentence-View* task as opposed to the *Abstract-View* task. The investigation leads us to the conclusion that employing sentence-level assessment achieves higher recall. This finding is also in alignment with previous studies (Rahbariasl & Smucker, 2019; Smucker & Jethani, 2010; Zhang et al., 2019).

Future research should extend the scope of the experiment to include participants with no medical background. This could test the hypothesis that, for the articles screening phase of the systematic review update process, a decrease in the level of knowledge of the assessor would not affect the assessment performance but would lead to a decrease in the associated costs. The main goal is to help individuals define their information needs with high precision, low burden, and minimal cost. Based on the participants' feedback, there appears to be an agreement on the usefulness of the computer-aided support in performing the task. Employing this system in other real-world cases, specifically where expert opinions are mandatory, could potentially speed up the process. The developed method could be implemented as a part of a IR system that aids medical experts in other tasks such as gathering test collections or biomedical articles curation. Another interesting future work would be adapting our system to other domains.

9 Conclusions

Motivated by the potential of Text Mining in unveiling hidden information from biomedical textual data, this dissertation has contributed to the Precision Medicine (PM) initiative using a Natural Language Processing (NLP) approach. At the beginning of this research we posed the following main research question:

MRQ — How can Biomedical NLP techniques support and advance the precision medicine approach through collection and analysis of clinical and medical textual resources?

In order to answer this question, this research followed the Design Science Research (DSR) framework of Hevner et al. Hevner et al. (2004) as illustrated in Figure 1.2 within Section 1.4. One of the guidelines within the DSR framework is to *design as an artifact* where research must produce a viable artifact such as a model, a method, or a construct. Chapters 2 and 3 of this dissertation have exploited NLP techniques to build deliverable frameworks that serve as computer-assisted solutions that address the shortcomings of existing systems. In Chapter 2 we present the *SNPcurator*, an online platform dedicated to assist biomedical experts in successfully extracting information from genomic studies. The platform enables automatic knowledge discovery in the genetics domain without relying on human annotations. In Chapter 3 we introduced the *PreMedOnto ontology*, enabling experts to better understand the relations and hierarchy of the concepts and terms related to the Precision Medicine domain.

Chapters 4-8 investigate, design, and evaluate models to detect inference in medical text. Natural Language Inference (NLI) models relations between snippets of texts by classifying them as *Entailment, Neutral* and *Contradiction.* Extracting conflicting evidence-based medicine outcomes has a high potential for supporting the PM approach by demonstrating that the traditional one-size-fits-all has inherent limitations. In Chapter 4 we applied traditional machine learning algorithms to the problem of detecting contradictions between two sentences extracted from abstracts of published articles. In Chapter 5 we extend the work to include two more inference relations, *Neutral* and *Entailment*, by re-purposing the same dataset to match the standard NLI benchmark format. The same chapter investigates the benefits of applying hybrid models through enriching deep networks with semantic features to boost the performance. In Chapter 6 we performed a thorough evaluation of state-of-the-art embedding models to find the most suitable technique for various biomedical tasks. Our evaluation toolkit, *MedSentEval*, includes 7 different techniques evaluated on 10 datasets that belong to 5 classification tasks. The findings show that recent approaches in modeling words according to the context they appear in are superior to other models. Also, effective use of these models requires pre-training on a medical language model, i.e., training on domain-specific medical text, before fine-tuning to the specific task. Chapter 7 subsequently builds upon this conclusion, by exploiting the context-dependent BERT embeddings in an ensemble classification environment on the MedNLI benchmark. These four chapters introduced scientifically validated artifacts by implementing techniques and evaluating them on established datasets. They demonstrate how text inference can grasp the meaning of the medical text, independent of differences in expression, within the medical domain.

Finally, in Chapter 8 we designed a user experiment that takes advantage of the prior NLI models to help experts in assessing the relevance of scientific articles. The experiment compares the precision levels and time frames when judging the articles by viewing abstracts as opposed to solely viewing the conclusion sentence and its inference label (e.g. entailment, neutral, or contradiction).

9.1 Contributions

In the introduction chapter, we posed seven research questions to investigate various aspects of the main research question. In this section we briefly review the contributions of each individual chapter to answer these research questions, and formulate a conclusion for each of them. Together, they constitute the answer to this dissertation's main research question.

RQ1 — How can text mining techniques be employed to extract relevant information from genome-wide associations studies?

In order to fill the gap between advances in the BioNLP research and the medical community, we have designed and implemented domain-specific NLP framework for fast content curation of genotype-phenotype related data from scientific literature. The web-based system automatically extracts Single Nucleotide Polymorphisms (SNPs) associations from genome-wide association studies. The design of the proposed framework takes into consideration the requirements and constraints set by expert curators in gold-standard databases. The information extracted includes SNP identifiers, source PubMed IDs that the entries belong to, the reported statistical significance parameters (P-value and OR), the population on which the study was conducted on, and the evidence sentence where the SNP mention occurred. The documents are retrieved by querying the PubMed repository, the query combines the user input and relevant genetic search terms. The proposed framework consists of a text mining pipeline that uses regular expressions to identity SNP mentions, follows a rule-based approach to couple the identified SNPs to their corresponding statistical significance and employs the SpaCy named entity recognition and dependency parser modules for extracting ethnicity and sample size of the cohort. The system interface allows users to visualize the extracted information in tabular format, an easy and straight-forward format to understand and analyze data. In the evaluation phase, two case studies were performed, where a disease term was queried through both *SNPcurator* and the NHGRI-EBI Catalog of manually-curated genome-wide association studies. Results show that *SNPcurator* was able to replicate a large number of SNP-disease associations that were also reported in the catalog. We further validate the performance of the extraction algorithm by on the *SNPPhenA*, a corpus of ranked associations of SNPs and phenotypes extracted from literature. We evaluated the performance of extracting a *(SNP, P-value)* pairs in terms of precision, recall and F-measure. The model achieved 81%, 86% and 83%, respectively, for each metric. Moreover, a qualitative evaluation of the usability and applicability of the tool was carried out by asking experts to try queries of their choice and answer a questionnaire. Notes and comments provided by the experts were taken into consideration, and an updated version of the tool is now deployed at `http://snpcurator.science.uu.nl`.

Conclusion I — Building a text mining pipeline using sentence splitting, sentence tokenization, regular expressions, name entity recognition and dependency parsing can lead to effective information extraction from unstructured scientific abstracts. Using the *SNPcurator* online information extraction platform, researchers and genetics' experts can automate the curation of SNPs and their corresponding information mentioned in GWA studies. While being limited to information included in the abstract text, the quantitative and qualitative results show great benefits and huge potential in automating or semi-automating the extraction of SNPs-trait information from literature. Accordingly, this proves useful in preliminary stages of genetics research where scientists need to conduct a full screening of the available literature.

RQ2 — How to structure knowledge extracted from large collections of scientific literature?

The development of standard organizations of knowledge and data of Precision Medicine is largely driven by the evolving needs of PM research that needs a shared hierarchy of its concepts. To better represent the knowledge on Precision Medicine, we have proposed an ontology learning pipeline based on the ontology reuse approach. Such an ontology should help experts in reaching a common understanding of the hierarchy and relations among the concepts related to the domain. The proposed framework consists of five stages; Knowledge Acquisition, Knowledge Formulation, Modular Reuse, Source Ontology Selection, and Ontology Enrichment. In the first two stages, we collect and process the data from two sources; a manually constructed vocabulary and scientific articles related to PM from the PubMed repository. At the Modular Reuse stage, we cluster the PM terms using the DBSCAN algorithm, according to their cosine similarity scores, where the top-ranked words per cluster serve as super-classes. In an ontology reuse framework, the quality and consistency of the newly developed ontology highly depends on the choice of source ontology. Therefore in the Ontology Selection stage, we use the NCBO bioportal recommender module to identify closely related gold-standard ontologies based for each cluster. According to the recommender, the following set of ontologies are the most relevant to the PM domain:

1. National Cancer Institute Thesaurus (NCIT)

2. Medical Subject Headings (MeSH)

3. Interlinking Ontology for Biological Concepts (IOBC)

The NCIT ontology had the greatest matching score among all related ontolgies. This is justified since the oncology domain has been at the forefront of the precision medicine revolution with the majority of all approved precision medicine diagnostics and treatments. In the final stage, Ontology Enrichment, we use Ontofox to distribute the remaining terms to each module. The Ontofox web tool takes as input sources terms (the terms extracted from literature), parent classes (the seed ontology modules) and the base ontology (the recommended ontology), and returns a hierarchy with intermediate classes between input child and parent with class annotations properties. Ontofox helps in selective class imports instead of importing the ontology as a whole.

For evaluation, We apply the Ontology Quality evaluation framework and Requirements framework (OQauRE) and the OntOlogy Pitfall Scanner (OOPS) to assess the *PreMedOnto* quality and correctness, respectively. OOPS validates ontologies against 41 pitfalls, the generated report shows that *PreMedOnto* is free from critical and important pitfalls which ensures its consistency. The OQauRE metrics provides more quantitative indicators of the ontology's quality with values ranging from 1 to 5 where 3 is the minimum score. *PreMedOnto* scored 3.5 for structure, 4.2 for compatibility and 4.5 for maintainability. The final ontology consists of 543 classes, it can be viewed and downloaded at `https://bioportal.bioontology.org/ontologies/PREMEDONTO` in OWL, RDF, or CSV.

Conclusion II — Given the intersection of the Precision Medicine with several medical domains, structuring and organising the data within the domain successfully relies on a maximisation of knowledge reuse to minimize the human factor and reduce redundancy. The proposed ontology learning process involves mining the PubMed repository to extract domain specific abstracts and vocabulary as sources of data. The information gathered is clustered and outlined to determine main modules. It takes advantage of existing knowledge bases by reusing and importing terms and concepts from published ontologies. *PreMedOnto*, built on top of gold-standards biological ontologies, provides a structured understanding of the general, investigations, diagnostics and treatment terms related to the Precision Medicine domain.

RQ3 — How to incorporate text mining methods in detecting contradictory statements found in scientific literature?

Abstracts of scientific literature may hold valuable information, as it sums up the research goals and findings. We designed and implemented a two-phase model to extract relevant conclusion sentences from abstract texts and classify whether it contradicts or entails the original hypothesis. We address the claim extraction process as a ranking problem, where a fixed-length feature vector is fed to a Learning to Rank algorithm. The algorithm uses a LambdaMART function to rank each input sentence according to its relevance to the query. At this stage, more insight into the medical context of the sentences is given by deploying word embeddings trained on a large collection of PubMed articles. In the second phase, the contradiction detection phase, we build on the assumption that negation is still the most useful feature of identifying oppositeness. For that reason, we employ NegEx to identify negation in textual medical data through regular expressions and predefined lists of trigger words. We also include antonyms as a feature to represent contradictions and add an alignment score to model entailment.

We evaluated our model on the *ManConCorpus* dataset consisting of published literature related to cardiovascular disease with a total of 259 articles consisting of 79 contradictory and 180 entailment claims. The dataset included structured and unstructured abstracts split into two sets for training and testing; the training set consisted of all abstracts with structured format (abstracts where the text is divided into subsections), while the test set included all unstructured abstracts (raw text). The claim extraction phase resulted in 0.97, 0.876, 0.91 average precision, recall, and F1, respectively. In the second phase, we were able to differentiate between contradictory and entailed claims with a precision of 0.9, a recall of 0.91, and an F1 score of 0.9.

Conclusion III — Detecting conflicts among published findings can be regarded as a two-phase detection model, namely, finding relevant sentences and detecting contradictions. Approaching the claim extraction problem using Learning to Rank algorithms is promising. The findings show that incorporating domain knowledge for extracting the most relevant sentence improves the final results. The proposed model was also able to improve the accuracy of detecting inference, namely the contradictory category, and still maintain good results regarding the entailment. The achieved improvement is mainly due to the enhanced negation detection through NegEx, and the inclusion of antonyms.

RQ4 — To what extent can deep learning improve the detection of textual inference, compared to traditional machine learning techniques?

To detect inference between texts, an important task in NLP, we found that models built for open-domain language could not be directly transferred to medical textual data. This is mainly due to the complexity and nature of the text that holds valuable medical knowledge. Given the unavailability of datasets, we enriched the same corpus used in the previous chapter to fit the standard NLI benchmarks; this required the addition of Neutral sentences that were not originally included. The new dataset consisted of 2135 records, with 1080, 608, and 447 entailment, contradiction, and neutral sentence pairs, respectively. We then used an experimental design to compare different machine learning techniques using cross-validation settings.

We experimented with classical models such as Naive Bayes, Support Vector Machines, Linear regression, Random Tree and a Gradient boost with a traditional feature vector of size 20. The features belong to three groups: String-Based Features, Contradiction-Based Features, and Context-Based Features that include word embeddings trained on PubMed articles. By surveying literature, one of the state-of-the-art techniques in detecting inference is the InferSent model trained on the SNLI dataset, a collection of 570k human-generated English sentence-pairs, and a bi-directional Long Short Term Memory (BiLSTM) encoder. We also experiment with the Universal Sentence Encoder (USE), which follows a Multi-task Learning architecture with data from different sources and partially augmented with instances from the SNLI corpus. Moreover, we designed a feature-assisted neural network that incorporated sentence embeddings and traditional features with a Dense Neural Network.

Regarding the effect of the feature-groups, the contradiction-based features had the most significant influence on the performance. The Random Forest and XGBoost algorithms achieved the best results among traditional ML algorithms, with the latter outperforming the former and achieving an accuracy of 76.94%. Of the two Deep Learning models, especially the InferSent shows excellent performance. However, there was no added accuracy when varying the number of layers or nodes. The InferSent model achieved a maximum accuracy of 93.95%, while the USE model achieved a maximum accuracy of 83.68%. Using the hybrid model generally results in higher performance, however, there is a variation in the performance gain between both models. There was a small improvement when incorporating the feature vector to the network with a difference of approximately 0.6% and 2.3% increase for the USE and InferSent model, respectively.

Conclusion IV — Deep learning techniques, both to represent and to classify text, give better performance improvement over traditional machine learning techniques when detecting language inference in medical text. Even with a modestly sized dataset, the results suggest that the traditional machine learning features and novel deep learning algorithms are complementary. Their combination in an end-to-end model enhanced the learning process and improved the predictions on unseen data.

RQ5 — Which textual representations are most suited for the biomedical domain?

The biomedical domain lacks a full and thorough analysis of sentence embedding techniques on common grounds, specifically with the fast developmental pace of neural embedding methods in recent years. For that purpose, we created *MedSentEval*, a python toolkit for evaluating state-of-the-art sentence embedding methods transferability to BioNLP tasks. The toolkit includes 7 different embedding models that belong to 3 categories: traditional word embeddings (GloVe and FastText models), contextualized embeddings (ELMo, BERT, and Flair models) and dedicated sentence encoders (InferSent and USE encoders). Apart from the sentence encoders category, we use the Mean-of-word-embeddings scheme to create sentence embedding by averaging single word vectors. We follow the original *SentEval* strategy in using simple classification models such as logistic regression or a single layer MLP on top of the generated sentence representations. This was motivated by the desire to assess the ability of each model independently, without the help of complex classification algorithms, in representing medical information. All models were trained and tested on 10 biomedical datasets with different objectives, sizes, and number of classes. The corresponding BioNLP tasks included semantic similarity, natural language inference, question answering, sentence classification, and sentiment analysis. Based on the results and evaluation over all datasets included in *MedSentEval*, we identified 5 key observations:

1. Models still need to resume training on domain-related and task-specific data.
2. Mean of word embeddings is still effective in capturing the sentence semantics compared to dedicated sentence encoders.
3. There is no correlation between the embedding vector size and the performance achieved.
4. Models based on language models usually outperform other encoding schemes.
5. A proper balance and variation in the training resources can lead to better performance.

The source codes and usage scenarios for reproducing results and/or extending the evaluation could be found at `https://github.com/nstawfik/MedSentEval`

Conclusion V — Context-dependent embedding models, in particular, ELMo and BERT, are superior to all other textual representation techniques. The quantitative results of the *MedSentEval* evaluation benchmark show there is still a room for improvement to reach a single universal embedding model for medical text that generalizes well over tasks. Given the obtained quantitative results, we recommend integrating biomedical language models as a promising direction of research given its ability to capture clinical and medical knowledge found in the input text compared to other methods.

RQ6 — What are the benefits of using state-of-the-art embeddings for recognizing clinical text entailment?

Whereas Chapter 6 established that context-dependent models are a promising approach in representing medical text, we have further investigatd this in Chapter 7 to determine the validity and generalizability of our findings. This chapter details our participation in two independent tasks of the MEDIQA challenge: the NLI and RQE tasks. The two tasks tackle the problem of inference among textual data while providing modest-size benchmarks for training, validation, and testing to overcome the lack of annotated resources common in the biomedical domain.

The NLI task aims to detect inference relations between clinical sentence pairs extracted from intensive care notes of the MIMIC database. The final NLI dataset consisted of 11,232, 1,395, and 1,422 for training, development, and testing, respectively. an additional test batch of 405 unlabelled pairs was provided by the challenge organizers. We submitted a total of 5 runs to the task achieving a maximum of 0.852 accuracy for the test set. Our best run ranked 12th over 143 submitted runs. The corresponding architecture used an ensemble classification with 3 stacking levels. The first level included 30 BERT base predictions using two pre-trained models (*SciBERT, ClinicalBERT*). A meta-learner was used for the second level by training a k-Nearest Neighbors, a Support-Vector Machine, and a Naive Bayes classifier. In the last level, we trained a logistic regression classifier on top of the combined predictions generated from the previous level.

The RQE task, again consisting of entailed and contradictory pairs, but among questions, aims at detecting duplicate questions asked by patients. Even though this work is more focused on the NLI task between informative sentences but given the similarity of both tasks, we attempted to solve the RQE task as well. The dataset included 8,588, 302, and 230 pairs for training, validation, and testing, respectively. Our model relied on the Universal Sentence Encoder (USE) embeddings fed to a logistic regression classifier. Even though our approach for this task was much simpler than the first task, we still managed to achieve a considerably good accuracy outperforming the baseline by 4.3%. Our team ranked 23rd among all 54 participants.

Conclusion VI — Detecting inference relations between pairs of sentences or questions with novel text representation models is possible, with an accuracy that outperforms original state-of-the-art results reported. In line with existing research, cluster ensembles are an effective method to increase the performance of single base models. In the NLI task, consistent improvement is obtained from successive ensembling from one to two stacking levels.

RQ7 — How can BioNLP models be employed to help domain experts in the information retrieval of evidence?

It is becoming more challenging for health professionals to keep up to date with current research, specifically with the adoption of Precision Medicine into practice as they have to extract extra information linking genomic variants to phenotypic data to make informed, personalized clinical decisions per patient. To save time and efforts, many medical experts prefer to curate systematic reviews instead of primary studies. Subsequently, there is a need to check reviews to keep them up-to-date with current studies. This final question addresses the evidence syntheses process of medical articles through the case study of updating systematic reviews. In that context, relevance assessment refers to the task of determining the relevance of an article with respect to a review topic. It can be interpreted as a measurement in which assessors judge the pertinence of a document to a chosen topic.

We designed and conducted a controlled user experiment where the primary outcome of interest is to compare the performance levels achieved when judging full abstracts versus single sentences accompanied by Natural Language Inference (NLI) labels. The experiment included 22 participants comprised mainly of academic professionals, with a high-level of expertise, familiar with the process of conducting a systematic review. A total of 143 PubMed articles were collected that belong to 6 different systematic reviews extracted from the Cochrane Database of Systematic Reviews (CSDR). This chapter partly builds on our previous models to generate the NLI labels. We pair the conclusion sentence of the original review with the finding sentence of each candidate study to find their inference relation. The experiment session comprised two main tasks carried out in sequence; namely, *Abstract-view* task and *Sentence-view* task with an imposed time limit of 15 minutes per task. It also included pre and post tasks questionnaires for demographics information and feedback survey. The experiment was created and hosted on the online research platform Gorilla (https://gorilla.sc/).

Participants achieved 0.56, 0.59, and 0.57 in the *Abstract-view* task and 0.66, 0.66, 0.64 in the *Sentence-view* task for accuracy, recall, and F2 respectively. The findings lead us to the conclusion that employing sentence-level for assessment achieves higher recall. The post-experiment questionnaire shows that half of the participants agree that time limits imposed extra stress on the judgment process with an inter-participant average agreement of 2.67 (on a scale from 1 to 5). The results show that participants felt more confident and believe their answers were more accurate in the Sentence-View task. The feedback on the usability of the NLI labels show that participants favored the computer-assisted suggestions, and 98% would recommend the system to a friend.

Conclusion VII — Following the ”human-in-the-loop” approach by engaging medical experts through a controlled user experiment is a useful methodology to demonstrate the utility afforded by BioNLP models and techniques. The case study shows that computer-assisted suggestions, in specific the natural language inference labels, can be used to support relevance assessment for updating systematic medical reviews. The accuracy and recall of relevance judgment by participants are higher when viewing less text and NLI labels rather than viewing full abstracts. This leads to a decrease in the associated costs and overall efforts of medical experts.

9.2 Research Validity

This section discusses the validity and limitations of the research in this dissertation.

English language — Throughout this thesis, we have explored textual data from scientific publications or patient records that were entirely based on English text. The PM domain, striving to find relations between phenotyes, genotypes and environmental factors among diverse individuals and populations should not be bound to a single language. This issue was not neglected by choice, but given the lack of annotated resources in the biomedical domain in general, medical or clinical datasets in other languages are very scarce and do not cover the topics of our work. In doing so, we disregard all knowledge contained in non-English resources languages. As a consequence, all data reported about patients in non-English speaking countries is severely under-analysed. While this is true for all types of data, the data from published literature is the least affected given that English is considered the lingua franca in science (Kamadjeu, 2019). On the other hand, the clinical notes or electronic health records that are the main sources of individual patient data remain strongly affected by the limitation of the language.

Data bias — In Chapter 4 we relied on the *ManConCorpus*, originally designed to detect conflicts between sentences with YES/NO responses to medical claims. In our first attempt to model inference in the biomedical domain, we used an extended version of the same corpus after enriching its content to match the format of NLI benchmarks. Chapter 5 details the repurposing process, but we highlight that the main modification was to include extra sentences that do not exist in the original corpus. We enriched the corpus by adding the first sentence of each abstract and assigning them with the label *NEUTRAL*. While human experts annotated the original corpus, our modified version is semi-annotated since the added phrases were not verified by a human. The first sentence usually describes the objective of the research, we argue that our choice for the objective sentence was not at random, but based on the observation of neutral sentences across different NLI benchmarks that are usually constructed by adding a purpose clause Gururangan et al. (2018).

In Chapter 7 the MEDIQA organizers provided an extra test set for the MedNLI task consisting of 405 instances. While the test set is relatively small, the main issue in the data is the label pattern. Any premise is always paired with three consecutive hypotheses denoting a repetitive class sequence (entailment, contradiction, and neutral). Such a design might be captured through the learning process and hence influence the performance of the classifier. In other words the model might memorise the pattern and classify accordingly. While we do not believe that this is the case for our model, this observation explains the high performance of other methods that are primarily due to the bias of this particular set.

Language Model Bias — When employing the contextualized word embeddings, we took advantage of the *ClincalBERT* model pre-trained on all note types in the MIMIC-III database, including nursing and physician notes, ECG reports, imaging reports, and discharge summaries. The MedNLI benchmark corpus also includes a small subset of the MIMIC notes. This could be referred to as the language model bias since, by using the pre-trained weights, we cannot remove individual notes that appeared in any of the train, development or, test sets. We believe that this bias has minimal or no effect given the dramatically larger size, approximately 2 million notes, of the entire MIMIC dataset as opposed to the MedNLI dataset. We have also tried to mitigate this issue by including a second BERT model, SciBERT, pre-trained on scientific literature in an attempt to further minimize the impact of this problem.

Expert validation — Some of the research artefacts that were described in this dissertation, namely in Chapter 2 and 8, collect data that captures the domain experts' perceptions. For that, we used surveys and questionnaires, on specific case-studies, as an exploratory method for eliciting and validating these artefacts. Although we have strived to select a random representative group of participants, the number of participants was usually modest, again leaving openings for shortcomings potentially left unreported, thus introducing a risk that our results may not represent all potentially relevant input yet.

9.3 Future Research

This dissertation has addressed several of the current challenges in applying text mining on biomedical textual data, yet there are still many possibilities to be explored and realized within this topic of research. Throughout the dissertation, we have shown that building and employing pipelines of existing natural language processing techniques to clinical notes or medical literature can indeed support the Precision Medicine revolution. With the continuous emergence of data, plenty of directions for future research can be imagined. We describe a few of them below.

Usage of full-text articles — One of the limitations of mining merely scientific abstracts is the potential loss of information that might be excluded from the abstract text, but is available in the full-text articles or even supplementary materials. Relying on abstracts is a common practice for text mining systems, mainly due to the accessibility of abstract texts via scientific databases and their corresponding APIs. However, analyzing full-text scientific publications allows for a broader information diversity, higher volume, and extracts secondary findings. In an IE framework like *SNPcurator.* this would be very promising since, logically, full-text contains more genomic entities and relations between those entities. However, while it opens the door for more information, additional steps are needed to clean and normalize the unstructured format of full-text articles, usually PDFs, and there is also more room for false positives.

Insights from Patients — At the beginning of this research, one of our long-term ambitions was to gain insights from patient-authored texts (PAT), to align with the original idea that motivated Precision Medicine in the first place: to put the patient at the center of healthcare. We hoped to collect and analyze ePATs, from different sources like Twitter, online health forums, or disease-specific portals, on the topic of PM and its related tests and treatments. Low levels of genomic and PM literacy hindered our hopes and ambitions; patients still do not realize the benefits offered by genomics-based technologies, nor understand how PM influences their care. (Literacy et al., 2016; Ramos, Ramos, & Ramos, 2019; Wynn et al., 2018). This could be attributed to the lack of educational material and the relatively complicated language when describing the concept. However, PM is gaining traction, from primary care to oncology, we expect that sooner than later, a wider population of patients will interact with PM results or get involved in issues regarding PM such as privacy concerns of human participants leading to more data for further analysis.

Building a Biomedical Benchmark — Research in biomedical NLP (BioNLP) is increasingly driven by the acquisition and construction of large datasets. One of our goals was to create a textual inference benchmark with sentence pairs extracted from published biomedical articles. As opposed to MEDNLI, the language used in clinical notes can be very different from biomedical literature and also sentences are relatively shorter. Moreover, the existing benchmark suffers from a patterned bias, as pointed out in section 9.2, and so more sets are needed to validate the state-of-the-art models. On the other hand, the dataset used to fit the textual entailment task from Chapter 5 is highly skewed towards the entailment class. We did not manage to construct an improved dataset due to time constraints. However, it is not hard to imagine several more challenges that arise when attempting to create a biomedical benchmark dataset. An example is the complex nature of biomedical literature that often necessitates domain-specific knowledge and expertise. As a consequence, building a corpus relies mainly on manual annotation by professionals associated with a high cost both in terms of time and money. Another dilemma is the lack of common models for manual and automatic annotations among the BioNLP community. Even when following general-domain annotation guidelines, there is often inconsistency in interpreting them among annotators. This results in low inter-annotator agreement rates leading to poor performances of models trained on these resources.

9.4 Personal Reflection

In these final paragraphs, I will take the opportunity to reflect on this research, and on some of the lessons I acquired during this journey, both on the academic and personal level.

Aim High, Dream Big — In the beginning of my PhD, I was constantly overwhelmed by several factors: I wanted to conduct quality research, my project had to be realistic and feasible within my timeline, I had to meet the publication requirements to graduate, all while carrying out my job as an assistant lecturer in my home university. Many days, I thought that this was an over-ambitious plan, and slowly, the imposter syndrome was kicking in, and I just felt inadequate with everything I do. I'm glad I kept going on and didn't give up as I realized that the only way to make it work was to believe in myself. I also realized that I am the one in control, and that all I had to do was to work hard to become the best version of myself. Before my PhD, I never thought I could learn so much, adapt to a new field so quickly, and step so far out of my comfort zone in order to reach an upcoming target. I was able to grow academically in many ways I never imagined before. The past 4 years are a reminder to always aim high, because even if I may dip a little, but I should never ever aim low.

Knowledge is of No Value Unless You Put it into Practice — I always wanted to conduct practical research, where my work would be directed towards solving problems facing the medical experts on a daily basis. But the truth is that in the academic world, your worth is defined by your publication and citation counts. While this is the most straight forward form of knowledge sharing, coping with the lengthy peer-review process and revision cycles, especially in a fast-evolving field like NLP, made me wonder if it is truly worth it. In the beginning of my degree, I invested a lot of time in building the *SNPcurator*, and even more so in getting this work published. However, ever since it was brought to light, I receive, every now and then, messages from biomedical researchers either inquiring about its usage, asking about extra features, or simply showing interest in the tool. While it might seem trivial, these messages always gave me power boosts to move forward. The feeling that one is making an impact, even if small, is utterly rewarding and a constant reminder of why I started this journey in the first place. My hope is that this dissertation has made a positive contribution, however small, to a future in which health care can even better fulfill the needs of caregivers and patients.

Life Does not Always Go As Planned — To keep track of my research goals, I was always keen on making plans and drafting to-do lists. However, the nature of research means that things might not always go according to these plans. No matter how good and precise they were, there were times where my research would get sidetracked because of a failed experiment, some unexpected results, persistent implementation errors, or simply getting caught in a never-ending reading cycle about topics that seem unrelated. I learned that in situations like these, I should keep calm, take a step back, reflect on things, and finally carry on. I also learned that abandoning my crafted plans and pursuing an accidental discovery could lead to very successful endings. Yet, during the last year of my PhD journey, the COVID-19 pandemic made it difficult to plan the next chapter of my life. Finalizing my dissertation, preparing for my defense that might not even happen because of the pandemic, felt overwhelming. My worries about "what will happen to my PhD now" turned into a personal challenge to make my thesis stronger than before. I managed to publish my 7^{th} paper, exceeding my pre-set publication goals. While the pandemic has delayed some aspects of my PhD and has altered my career plans, personally, I'm grateful for the time I took off work-related stress and pre-determined deadlines.

Bibliography

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *5th international conference on learning representations, {iclr} 2017, toulon, france, april 24-26, 2017, conference track proceedings.* Retrieved from https://openreview.net/forum?id=BJh6Ztuxl

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data.* Springer {US}. doi: 10.1007/978-1-4614-3223-4

Agyeman, A. A., & Ofori-Asenso, R. (2015). Perspective: Does personalized medicine hold the future for medicine? *Journal of pharmacy & bioallied sciences*, *7*(3), 239.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649. Retrieved from https://aclanthology.info/papers/C18-1139/c18-1139

Alamri, A. (2016a). *The detection of contradictory claims in biomedical abstracts* (Unpublished doctoral dissertation). University of Sheffield.

Alamri, A. (2016b). *The Detection of Contradictory Claims in Biomedical Abstracts* (Unpublished doctoral dissertation). University of Sheffield.

Alamri, A., & Stevenson, M. (2015). Automatic detection of answers to research questions from medline abstracts. In *Proceedings of bionlp* (Vol. 15, pp. 141–146).

Alamri, A., & Stevenson, M. (2016a). A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, *7*(1), 36.

Alamri, A., & Stevenson, M. (2016b, 6). A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, *7*, 36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27267226http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4897929 doi: 10.1186/s13326-016-0083-z

Alamri, A., & Stevensony, M. (2015, November). Automatic identification of potentially contradictory claims to support systematic reviews. In *Proc. ieee int. conf. bioinformatics and biomedicine (bibm)* (pp. 930–937). doi: 10.1109/BIBM.2015.7359808

Ali-Khan, S., Kowal, S., Luth, W., Gold, R., & Bubela, T. (2016). *Terminology for Personalized Medicine: a systematic collection Terminology for personalized medicine* (Tech. Rep.).

Alobaidi, M., Malik, K. M., & Hussain, M. (2018, 10). Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain. *Computer Methods and Programs in Biomedicine*, *165*, 117–128. doi: 10.1016/j.cmpb.2018.08.010

Alobaidi, M., Malik, K. M., & Sabra, S. (2018, 12). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics*, *19*(1), 319. doi: 10.1186/s12859-018-2339-3

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019, 4). Publicly Available Clinical BERT Embeddings. *arXiv e-prints*, arXiv:1904.03323. Retrieved from http://arxiv.org/abs/1904.03323

Amato, F., Santo, A. D., Moscato, V., Picariello, A., Serpico, D., & Sperli, G. (2015, 7). A Lexicon-Grammar Based Methodology for Ontology Population for e-Health Applications. In *2015 ninth international conference on complex, intelligent, and software intensive systems* (pp. 521–526). IEEE. doi: 10.1109/CISIS.2015.76

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*. doi: 10.3758/s13428-019-01237-x

Arguello Casteleiro, M., Demetriou, G., Read, W., Fernandez Prieto, M. J., Maroto, N., Maseda Fernandez, D., ... Stevens, R. (2018, 12). Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of Biomedical Semantics*, *9*(1), 13. doi: 10.1186/s13326-018-0181-1

Arora, S., Liang, Y., & Ma, T. (2017). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *Proceedings of the international conference on learning representations.* Retrieved from https://pdfs.semanticscholar.org/3fc9/7768dc0b36449ec377d6a4cad8827908d5b4.pdf

Bashir, R., Surian, D., & Dunn, A. G. (2018, 12). Time-to-update of systematic reviews relative to the availability of new evidence. *Systematic Reviews*, *7*(1), 195. Retrieved from https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-018-0856-9 doi: 10.1186/s13643-018-0856-9

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., & Brookes, A. J. (2014, 7). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum*

Genet, *22*(7), 949–952. Retrieved from http://dx.doi.org/10.1038/ejhg.2013.274 doi: 10.1038/ejhg.2013.274

Beltagy, I., Cohan, A., & Lo, K. (2019). *SCIBERT: Pretrained Contextualized Embeddings for Scientific Text* (Tech. Rep.). Retrieved from https://github.com/huggingface/pytorch-pretrained-b

Ben Abacha, A., & Demner-Fushman, D. (2016). Recognizing Question Entailment for Medical Question Answering. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2016*, 310–318. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28269825http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333286

Ben Abacha, A., Shivade, C., & Demner-Fushman, D. (2019a). Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In *Acl-bionlp.* Florence.

Ben Abacha, A., Shivade, C., & Demner-Fushman, D. (2019b, 9). Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In (pp. 370–379). Association for Computational Linguistics (ACL). doi: 10.18653/v1/w19-5039

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., Ca, J. U., Kandola, J., . . . Shawe-Taylor, J. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155. Retrieved from http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Biomedical Semantic Question Answers. (2018). Retrieved March 3, 2019. Retrieved from http://www.bioasq.org/participate/useful_resources

Biomedical Semantic Similarity Estimation System. (2017). Retrieved February 28, 2019. Retrieved from http://tabilab.cmpe.boun.edu.tr/BIOSSES/DataSet.html

Boca, S. M., Panagiotou, O. A., Rao, S., McGarvey, P. B., & Madhavan, S. (2018, 4). Future of Evidence Synthesis in Precision Oncology: Between Systematic Reviews and Biocuration. *JCO Precision Oncology*(2), 1. doi: 10.1200/po.17.00175

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017, 12). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from https://www.aclweb.org/anthology/Q17-1010 doi: 10.1162/tacl{_}a{_}00051

Bokharaeian, B., Diaz, A., Taghizadeh, N., Chitsaz, H., & Chavoshinejad, R. (2017). SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature. *Journal of biomedical semantics*, *8*(1), 14.

Bontas, E. P., Mochol, M., & Tolksdorf, R. (2005). Case Studies on Ontology Reuse. In *In proc. of the 5th international conference on knowledge management.*

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing (emnlp).* Lisbon: Association for Computational Linguistics.

Bravo, , Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015, 2). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *{BMC} Bioinformatics*, *16*(1). Retrieved from `http://dx.doi.org/10.1186/s12859-015-0472-9` doi: 10.1186/s12859-015-0472-9

Bruno, A. E., Li, L., Kalabus, J. L., Pan, Y., Yu, A., & Hu, Z. (2012). {miRdSNP}: a database of disease-associated {SNPs} and {microRNA} target sites on 3{\textquotesingle}{UTRs} of human genes. *{BMC} Genomics*, *13*(1), 44. Retrieved from `http://dx.doi.org/10.1186/1471-2164-13-44` doi: 10.1186/1471-2164-13-44

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology Learning from Text: Methods, Evaluation And Applications.
doi: 10.1162/coli.2006.32.4.569

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, *11*(23-581), 81.

Cahyani, D. E., & Wasito, I. (2017, 6). Automatic Ontology Construction Using Text Corpora and Ontology Design Patterns (ODPs) in Alzheimer's Disease. *Jurnal Ilmu Komputer dan Informasi*, *10*(2), 59. doi: 10.21609/jiki.v10i2.374

Carlson, B. (2008). SNPs-A shortcut to personalized medicine. *Genetic Engineering & Biotechnology News*, *28*(12), 12.

Catillon, M. (2017a). Medical Knowledge Synthesis: A Brief Overview. Retrieved from `https://www.hbs.edu/faculty/Pages/item.aspx?num=54337`

Catillon, M. (2017b). Medical Knowledge Synthesis: A Brief Overview. Retrieved from `https://www.hbs.edu/faculty/Pages/item.aspx?num=54337`

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... Kurzweil, R. (2018, 3). Universal Sentence Encoder. *arXiv preprint*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St John, R., ... Kurzweil Google Research Mountain View, R. (2018). Universal Sentence Encoder. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174).

Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D18-2029

Chapman, W. W. (2010). Closing the gap between NLP research and clinical practice. *Methods of information in medicine*, *49*(4), 317–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20686731

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, *34*(5), 301–310.

Chaussabel, D. (2004). Biomedical literature mining: Challenges and solutions in the 'omics' era. *American Journal of PharmacoGenomics*, *4*(6), 383–393. doi: 10.2165/00129785-200404060-00005

Chawla, N. V., & Davis, D. A. (2013, 6). Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *Journal of General Internal Medicine*, *28*(S3), 660–665. doi: 10.1007/s11606-013-2455-8

Chen, Q., Kim, S., Wilbur, W. J., Du, J., & Lu, Z. (2018). Combining rich features and deep learning for finding similar sentences in electronic medical records. In *Proceedings of the biocreative/ohnlp challenge, 2018.*

Chen, Q., Peng, Y., & Lu, Z. (2018). *BioSentVec: creating sentence embeddings for biomedical texts* (Tech. Rep.). Retrieved from https://github.com/ncbi-nlp/BioSentVec.

Chen, R.-C., Spina, D., Croft, W. B., Sanderson, M., & Scholer, F. (2015). Harnessing semantics for answer sentence retrieval. In *Proceedings of the eighth workshop on exploiting semantic annotations in information retrieval* (pp. 21–27).

Chen, R.-C., Yulianti, E., Sanderson, M., & Bruce Croo, W. (2017). On the Benefit of Incorporating External Features in a Neural Architecture for Answer Sentence Selection. *ACM Reference format*. doi: 10.1145/3077136.3080705

Chen, Z., He, Z., Liu, X., & Bian, J. (2018, 7). Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC Medical Informatics and Decision Making*, *18*(S2), 65. Retrieved from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0630-x doi: 10.1186/s12911-018-0630-x

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to Train Good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing* (pp. 166–174). Berlin, Germany. Retrieved from https://www.aclweb.org/anthology/W16-2922 doi: 10.18653/v1/W16-2922

Chklovski, T., & Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Emnlp* (Vol. 4, pp. 33–40).

Choi, J. D., Tetreault, J., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing, acl* (pp. 26–31).

Chow, N., Gallo, L., & Busse, J. W. (2018, 9). Evidence-based medicine and precision medicine: Complementary approaches to clinical decision-making. *Precision Clinical Medicine*, *1*(2), 60–64. Retrieved from `https://academic.oup.com/pcm/article/1/2/60/5075444` doi: 10.1093/pcmedi/pby009

Citation Sentiment Analysis Dataset. (2015). Retrieved February 28, 2019 through personal communication.

Clinical Research Informatics — AMIA. (2019). Retrieved from `https://www.amia.org/applications-informatics/clinical-research-informatics`

Clinical Semantic Textual Similarity Dataset. (2018). Retrieved February 28, 2019 through personal communication.

Cohen, A. M., Ambert, K., & McDonagh, M. (2010). A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *Amia annual symposium proceedings* (Vol. 2010, p. 121).

Cohen, K. B., & Demner-Fushman, D. (2014). *Biomedical Natural Language Processing.* John Benjamins Publishing Company. Retrieved from `https://books.google.com.eg/books?hl=fr&lr=&id=vXvPAgAAQBAJ&oi=fnd&pg=PR1&dq=biomedical+natural+language+processing&ots=ZGbkQ5ApKO&sig=fMZ2Hic2iSeh8tqMS1e1-MdfoqE&redir_esc=y#v=onepage&q=biomedicalnaturallanguageprocessing&f=false`

Collier, N., Nazarenko, A., Baud, R., & Ruch, P. (2006, 6). Recent advances in natural language processing for biomedical applications. *International Journal of Medical Informatics*, *75*(6), 413–417. doi: 10.1016/j.ijmedinf.2005.06.008

Conneau, A., & Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).* Retrieved from `https://aclanthology.info/papers/L18-1269/l18-1269`

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017, 5).

Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv preprint*. Retrieved from http://arxiv.org/abs/1705.02364

A Corpus of Contradictory Research Claims from Cardiovascular Research Abstracts. (2016). Retrieved March 11, 2019..

Cowie, J., & Wilks, Y. (2000). Information extraction. In *Handbook of natural language processing* (p. 57). Marcel Dekker New York, NY. Retrieved from https://books.google.com.eg/books?hl=fr&lr=&id=VoOLvxyX0BUC&oi=fnd&pg=PA241&dq=+information+extraction+extract+specific+parts&ots=wvfYLJ5PoZ&sig=A_-0qePI-a50VJme4OZrmKNOTOs&redir_esc=y#v=onepage&q=informationextractionextractspecificparts&f=false

Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, *15*(4). Retrieved from http://nlp.uned.es/clef-qa/ave/ doi: 10.1017/S1351324909990209

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013a, 7). Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, *6*(4), 1–220. Retrieved from http://www.morganclaypool.com/doi/abs/10.2200/S00509ED1V01Y201305HLT023 doi: 10.2200/S00509ED1V01Y201305HLT023

Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. (2013b, 7). Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, *6*(4), 1–220. Retrieved from http://www.morganclaypool.com/doi/abs/10.2200/S00509ED1V01Y201305HLT023 doi: 10.2200/S00509ED1V01Y201305HLT023

Davis, F. D. (1989, 9). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. {*MIS*} *Quarterly*, *13*(3), 319. Retrieved from http://dx.doi.org/10.2307/249008 doi: 10.2307/249008

Del Corro, L., & Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on world wide web* (pp. 355–366).

De Marneffe, M.-C., Rafferty, A. N., & Manning, C. D. (2008). Finding contradictions in text. In *Acl* (Vol. 8, pp. 1039–1047).

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920–1930. Retrieved from https://www-ncbi-nlm-nih-gov.gate2.inist.fr/pmc/articles/PMC5831252/pdf/nihms729905.pdf doi: 10.1161/

CIRCULATIONAHA.115.001593

Dernoncourt, F., & Lee, J. Y. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the eighth international joint conference on natural language processing.* Taipei. Retrieved from `https://www.ncbi.nlm.`

de Silva, N., Dou, D., & Huang, J. (2017a). Discovering inconsistencies in pubmed abstracts through ontology-based information extraction. In *Acm conference on bioinformatics, computational biology, and health informatics (acm bcb), p. to appear.*

de Silva, N., Dou, D., & Huang, J. (2017b). Discovering inconsistencies in pubmed abstracts through ontology-based information extraction. In *Acm conference on bioinformatics, computational biology, and health informatics (acm bcb).*

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints.* Retrieved from `https://github.com/tensorflow/tensor2tensor`

Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T., & Kann, M. G. (2011, 2). Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. *Bioinformatics*, *27*(3), 408–415. Retrieved from `http://dx.doi.org/10.1093/bioinformatics/btq667` doi: 10.1093/bioinformatics/btq667

Douglas Zhang, X. (2015). Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. *J Pharmacogenomics Pharmacoproteomics*, *6*, 144. Retrieved from `http://dx.doi.org/10.4172/2153-0645.1000e144` doi: 10.4172/2153-0645.1000e144

Dramé, K., Diallo, G., Delva, F., Dartigues, J. F., Mouillet, E., Salamon, R., & Mougin, F. (2014, 4). Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application to Alzheimer's disease. *Journal of Biomedical Informatics*, *48*, 171–182. doi: 10.1016/J.JBI.2013.12.013

Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017, 12). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics*, *8*(1), 9. Retrieved from `http://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0120-6` doi: 10.1186/s13326-017-0120-6

Duque-ramos, A., Duque-ramos, A., Fernández-breis, J. T., Stevens, R., & Aussenac-gilles, N. (2011). OQuaRE: A SQuaRE-based approach for evaluating the quality of ontologies. *JOURNAL OF RESEARCH AND*

PRACTICE IN INFORMATION TECHNOLOGY, 159.

Elragal, A., & Haddara, M. (2019, 5). Design Science Research: Evaluation in the Lens of Big Data Analytics. *Systems*, *7*(2), 27. Retrieved from `https://www.mdpi.com/2079-8954/7/2/27` doi: 10.3390/systems7020027

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H., & Rosenbaum, M. E. (2005). Answering physicians' clinical questions: Obstacles and potential solutions. *Journal of the American Medical Informatics Association*, *12*(2), 217–224. doi: 10.1197/jamia.M1608

Fridsma, D. B. (2016, 11). Health informatics: our domain, our challenge. *Journal of the American Medical Informatics Association*, *23*(6), 1202–1202. Retrieved from `https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocw146` doi: 10.1093/jamia/ocw146

Gaizauskas, R., & Wilks, Y. (1998). *Information Extraction: Beyond Document Retrieval* (Vol. 3; Tech. Rep. No. 2).

Gao, M., Chen, F., & Wang, R. (2018). Improving Medical Ontology Based on Word Embedding.
doi: 10.1145/3194480.3194490

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., ... Zettlemoyer, L. (2018). AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of workshop for nlp open source software (nlp-oss)* (pp. 1–6). Melbourne: Association for Computational Linguistics. Retrieved from `https://aclweb.org/anthology/papers/W/W18/W18-2501/`

Gedzelman, S., Simonet, M., Bernhard, D., Diallo, G., & Palmer, P. (2005). Building an ontology of cardio-vascular diseases for concept-based information retrieval. In *Computers in cardiology, 2005* (pp. 255–258). IEEE. doi: 10.1109/CIC.2005.1588085

Ghoulam, A., Barigou, F., & Belalem, G. (2015, 4). Information Extraction in the Medical Domain. *Journal of Information Technology Research*, *8*(2), 1–15. Retrieved from `http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jitr.2015040101` doi: 10.4018/jitr.2015040101

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*, 337–355. Retrieved from `https://www.jstor.org/stable/43825912` doi: 10.2307/43825912

Gruber, T. R. (1993, 6). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. doi: 10.1006/KNAC.1993.1008

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018, 5). Annotation Artifacts in Natural Language Inference Data. In (pp. 107–112). Association for Computational Linguistics (ACL). doi: 10.18653/v1/n18-2017

Hao, Y., Liu, X., Wu, J., & Lv, P. (2018). Exploiting Sentence Embedding for Medical Question Answering. *arXiv e-prints*. Retrieved from `http://arxiv.org/abs/1811.06156`

Harabagiu, S., Hickl, A., & Lacatusu, F. (2006). Negation, contrast and contradiction in text processing. In *Aaai* (Vol. 6, pp. 755–762).

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2).

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004, 3). Design science in information systems research. *MIS Quarterly: Management Information Systems*, *28*(1), 75–105. doi: 10.2307/25148625

Hollenstein, N., de la Torre, A., Langer, N., & Zhang, C. (2019, 11). CogniVal: A Framework for Cognitive Word Embedding Evaluation. In (pp. 538–549). Association for Computational Linguistics (ACL). doi: 10.18653/v1/k19-1050

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 328–339). Melbourne: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P18-1031`

HPV Vaccination's Tweets Dataset. (2017). Retrieved March 9, 2019. Retrieved from `https://sbmi.uth.edu/ontology/files/TweetsAnnotationResults.zip`

Ikegawa, S. (2012). A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going. *Genomics & Informatics*, *10*(4), 220. doi: 10.5808/gi.2012.10.4.220

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannis Pavlopoulos, I. A., Aris Kosmopoulos. (2014). *Continuous space word vectors obtained by applying word2vec to abstracts of biomedical articles* (Tech. Rep.). NLP Group, Department of Informatics, Athens University of Economics and Business, Greece Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece.

Islamaj Dogan, R., Chatr-aryamontri, A., Kim, S., Wei, C.-H., Peng, Y., Comeau, D., & Lu, Z. (2017, 7). BioCreative VI Precision Medicine

Track: creating a training corpus for mining protein-protein interactions affected by mutations. In (pp. 171–175). Association for Computational Linguistics (ACL). doi: 10.18653/v1/w17-2321

Jameson, J. L., & Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *Obstetrical & gynecological survey*, *70*(10), 612–614.

Janes, J. W. (1994, 4). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. *Journal of the American Society for Information Science*, *45*(3), 160–171. Retrieved from `http://doi.wiley.com/10.1002/%28SICI%291097-4571%28199404%2945%3A3%3C160%3A%3AAID-ASI6%3E3.0.CO%3B2-4` doi: 10.1002/(SICI)1097-4571(199404)45:3⟨160::AID-ASI6⟩3.0.CO;2-4

Järvelin, K., & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international acm sigir conference on research and development in information retrieval* (pp. 41–48).

Jiménez-Ruiz, E., Cuenca Grau, B., Sattler, U., Schneider, T., & Berlanga, R. (2008). Safe and Economic Re-Use of Ontologies: A Logic-Based Methodology and Tool Support. In *Proceedings of the 5th european semantic web conference* (pp. 185–199). Tenerife.

Jin, D., & Szolovits, P. (2018). *PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks* (Tech. Rep.). Retrieved from `http://www.aclweb.org/anthology/W18-2308`

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., ... Mark, R. G. (2016, 5). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, 160035. Retrieved from `http://www.nature.com/articles/sdata201635` doi: 10.1038/sdata.2016.35

Kamadjeu, R. (2019, 9). English: the lingua franca of scientific research. *The Lancet Global Health*, *7*(9), e1174. doi: 10.1016/s2214-109x(19)30258-x

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition*. Springer International Publishing. doi: 10.1007/978-3-030-14596-5

Kang, Y., Fink, J. C., Doerfler, R., & Zhou, L. (2018, 10). Disease Specific Ontology of Adverse Events: Ontology extension and adaptation for Chronic Kidney Disease. *Computers in Biology and Medicine*, *101*, 210–217. doi: 10.1016/J.COMPBIOMED.2018.08.024

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). CLEF 2017 technologically assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–29).

Kanoulas, E., Li, D., Azzopardi, L., Spijker, R., & others. (2018). CLEF

2019 technology assisted reviews in empirical medicine overview. In *Clef 2018 evaluation labs and workshop: Online working notes. ceur-ws, france* (pp. 1–20).

Kelly, D., & Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends R in Information Retrieval*, *3*(2), 1–224. doi: 10.1561/1500000012

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1746–1751). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/D14-1181` doi: 10.3115/v1/D14-1181

Kiros, J., & Chan, W. (2018). InferLite: Simple Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4868–4874). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/D18-1524` doi: 10.18653/v1/D18-1524

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-Thought Vectors. In *Proceedings of the 28th international conference on neural information processing systems* (pp. 3294–3302). Montreal, Canada. Retrieved from `https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf`

Kitts, A., & Sherry, S. (2002). The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. *The NCBI Handbook. McEntyre J, Ostell J, eds. Bethesda, MD: US National Center for Biotechnology Information.*

Klinger, R., Furlong, L. I., Friedrich, C. M., Mevissen, H. T., Fluck, J., Sanz, F., & Hofmann-Apitius, M. (2007). *Identifying Gene Specific Variations in Biomedical Text.*

Koopman, B., & Zuccon, G. (2014). Why assessing relevance in medical IR is demanding. In *Proceedings of the sigir workshop on medical information retrieval (medir 2014).* Gold Coast.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019, 1). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Retrieved from `http://arxiv.org/abs/1901.08746`

Leslie, R., O\textquotesingleDonnell, C. J., & Johnson, A. D. (2014, 6). {GRASP}: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, *30*(12), i185–i194. Retrieved from `http://dx.doi.org/10`

.1093/bioinformatics/btu273 doi: 10.1093/bioinformatics/btu273

Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi: 10.1080/10447319509526110

Li, H., Caragea, D., Li, X., & Caragea, C. (2018). Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks. In *Proceedings of the iscram asian pacific 2018 conference.*

Li, M. J., Liu, Z., Wang, P., Wong, M. P., Nelson, M. R., Kocher, J.-P. A., ... Wang, J. (2015, 11). {GWASdb} v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res*, *44*(D1), D869–D876. Retrieved from http://dx.doi.org/10.1093/nar/gkv1317 doi: 10.1093/nar/gkv1317

Literacy, R. o. H., Practice, B. o. P. H., Health, P., Division, H., Medicine, National Academies of Sciences, E., & Medicine. (2016, 4). Communicating with the Public.

Liu, T.-Y., et al. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, *3*(3), 225–331.

Liu, Y., Liang, Y., & Wishart, D. (2015, 4). {PolySearch}2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Res*, *43*(W1), W535–W542. Retrieved from http://dx.doi.org/10.1093/nar/gkv383 doi: 10.1093/nar/gkv383

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). A Way to Automatically Enrich Biomedical Ontologies. In *Proceedings of the 19th international conference on extending database technology.* doi: 10.5441/002/edbt.2016.82

Lourenço, A., Carneiro, S., Carreira, R., Rocha, M., Rocha, I., & Ferreira, E. (2008). *A Tool for the Automatic and Manual Annotation of Biomedical Documents* (Tech. Rep.). Retrieved from http://www.ebi.ac.uk/chebi/

Mackin, R. (1978). On collocations: Words shall be known by the company they keep. In *In p. strevens (ed.), in honor of a. s. hornby. london* (pp. 149–165). Oxford: Oxford University Press.

Maddalena, E., Basaldella, M., De Nart, D., Degl'innocenti, D., Mizzaro, S., & Demartini, G. (2016). Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In *Fourth aaai conference on human computation and crowdsourcing.* Austin, Texas. Retrieved from www.aaai.org

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. doi: 10.1016/0167-9236(94)00041-2

McGowan, J., & Sampson, M. (2005, 1). Systematic reviews need systematic searchers. *Journal of the Medical Library Association*, *93*(1), 74–80.

The Medical Question Entailment Data. (2016). Retrieved March 3, 2019. Retrieved from `https://github.com/abachaa/RQE_Data_AMIA2016`

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th signll conference on computational natural language learning* (pp. 51–61). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/K16-1006` doi: 10.18653/v1/K16-1006

Metzler, D., & Kanungo, T. (2008). Machine learned sentence selection strategies for query-biased summarization. In *Sigir learning to rank workshop* (pp. 40–47).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Elsevier Inc. doi: 10.1016/C2010-0-66188-8

Mizzaro, S. (1997, 9). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832. Retrieved from `http://doi.wiley.com/10.1002/%28SICI%291097-4571%28199709%2948%3A9%3C810%3A%3AAID-ASI6%3E3.0.CO%3B2-U` doi: 10.1002/(SICI)1097-4571(199709)48:9⟨810::AID-ASI6⟩3.0.CO;2-U

Morton, S., Berg, A., Levit, L., Eden, J., & others. (2011). *Finding what works in health care: standards for systematic reviews.* National Academies Press.

Myles, S., Davison, D., Barrett, J., Stoneking, M., & Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics*, *1*, 22. Retrieved from `http://dx.doi.org/10.1186/1755-8794-1-22` doi: 10.1186/1755-8794-1-22

Naderi, N., & Witte, R. (2012). Automated extraction and semantic analysis of mutation impacts from the biomedical literature. {*BMC*} *Genomics*, *13*(Suppl 4), S10. Retrieved from `http://dx.doi.org/10.1186/1471-2164-13-S4-S10` doi: 10.1186/1471-2164-13-s4-s10

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011, 9). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. Retrieved from `https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000464` doi: 10.1136/amiajnl-2011-000464

A Natural Language Inference Dataset For The Clinical Domain. (2018). Retrieved March 4, 2019. Retrieved from `https://physionet.org/content/mednli/1.0.0/`

Newman-Griffis, D., Lai, A. M., & Fosler-Lussier, E. (2017). Insights into Analogy Completion from the Biomedical Domain. In *Proceedings of the 16th workshop on biomedical natural language processing (bionlp)* (pp. 19–28). Retrieved from `https://github.com/OSU-slatelab/BMASS.`

Niu, Y., Zhu, X., Li, J., & Hirst, G. (2005). Analysis of polarity information in medical text. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2005*, 570–4.

Norvig, P. (1987). Inference In Text Understanding. In *Aaai* (pp. 561–565). Retrieved from `www.aaai.org`

Obesity and overweight Fact Sheet. (2017). Retrieved from `http://www.who.int/mediacentre/factsheets/fs311/en/`

Ochs, C., Perl, Y., Geller, J., Arabandi, S., Tudorache, T., & Musen, M. A. (2017, 7). An empirical analysis of ontology reuse in BioPortal. *Journal of Biomedical Informatics*, *71*, 165–177. doi: 10.1016/J.JBI.2017.05.021

Oquendo, M. A., Baca-Garcia, E., Artés-Rodr\'\iguez, A., Perez-Cruz, F., Galfalvy, H. C., Blasco-Fontecilla, H., ... Duan, N. (2012, 1). Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry*, *17*(10), 956–959. doi: 10.1038/mp.2011.173

Padó, S., de Marneffe, M.-C., MacCartney, B., Rafferty, A. N., Yeh, E., & Manning, C. D. (2008). Deciding entailment and contradiction with stochastic and edit distance-based alignment. In *Tac.*

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 528–540). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/N18-1049` doi: 10.18653/v1/N18-1049

Panagiotou, O. A., Ioannidis, J. P. A., & others. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International journal of epidemiology*, *41*(1), 273–286.

Pazienza, M. T. (1999). *Information Extraction - Towards Scalable, Adaptable Systems.* Springer-Verlag Berlin Heidelberg. Retrieved from `https://www.springer.com/de/book/9783540666257` doi: 10.1007/3-540-48089-7

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011a). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011b). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D14-1162` doi: 10.3115/v1/D14-1162

Perone, C. S., Silveira, R., & Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv e-prints*. Retrieved from `https://arxiv.org/pdf/1806.06259.pdf`

Peters, M., Ruder, S., & Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *arXiv e-prints*. Retrieved from `https://arxiv.org/pdf/1903.05987.pdf`

Peters, M. E., Neumann, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. 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. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/N18-1202`

Pieper, D., Antoine, S. L., Neugebauer, E. A., & Eikermann, M. (2014, 12). Up-to-dateness of reviews is often neglected in overviews: A systematic review. *Journal of Clinical Epidemiology*, *67*(12), 1302–1308. doi: 10.1016/j.jclinepi.2014.08.008

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., . . . Furlong, L. I. (2015, 4). {DisGeNET}: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, *2015*(0), bav028–bav028. Retrieved from `http://dx.doi.org/10.1093/database/bav028` doi: 10.1093/database/bav028

Ponto, A. A.-B. A. J., PhD. (2015, 4). Understanding and Evaluating Survey Research. *Journal of the Advanced Practitioner in Oncology*, *6*(2). doi: 10.6004/jadpro.2015.6.2.9

Porter, M. F. (1980). An algorithm for suffix stripping.

Poveda-Villalón, M., Carmen Suárez-Figueroa, M., Ángel García-Delgado, M., & Gómez-Pérez, A. (2009). *OOPS! (OntOlogy Pitfall Scanner!): supporting ontology evaluation on-line* (Vol. 1; Tech. Rep.).

Prasad, V., Cifu, A., & Ioannidis, J. P. (2012a). Reversals of established medical practices: evidence to abandon ship. *Jama*, *307*(1), 37–38.

Prasad, V., Cifu, A., & Ioannidis, J. P. A. (2012b, 1). Reversals of established medical practices: Evidence to abandon ship. *JAMA - Journal of the American Medical Association*, *307*(1), 37–38. doi: 10.1001/jama.2011.1960

Prasad, V., Vandross, A., Toomey, C., Cheung, M., Rho, J., Quinn, S., ... others (2013). A decade of reversal: an analysis of 146 contradicted medical practices. In *Mayo clinic proceedings* (Vol. 88, pp. 790–798).

Preum, S. M., Mondol, A. S., Ma, M., Wang, H., & Stankovic, J. A. (2017a, 12). Preclude2 : Personalized conflict detection in heterogeneous health applications. *Pervasive and Mobile Computing*, *42*, 226–247. doi: 10.1016/J.PMCJ.2017.09.008

Preum, S. M., Mondol, A. S., Ma, M., Wang, H., & Stankovic, J. A. (2017b). Preclude: Conflict detection in textual health advice. In *Pervasive computing and communications (percom), 2017 ieee international conference on* (pp. 286–296).

PubMed 200k RCT Dataset. (2017). Retrieved March 3, 2019. Retrieved from `https://github.com/Franck-Dernoncourt/pubmed-rct`

PubMed PICO Element Detection Dataset. (2018). Retrieved March 5, 2019. Retrieved from `https://github.com/jind11/PubMed-PICO-Detection`

Rahbariasl, S., & Smucker, M. D. (2019, 7). Time-limits and summaries for faster relevance assessing. In *Sigir 2019 - proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 901–904). Association for Computing Machinery, Inc. doi: 10.1145/3331184.3331270

Rai, A. (2017, 3). Editor's Comments: Diversity of Design Science Research. *Management Information Systems Quarterly*, *41*(1). Retrieved from `https://aisel.aisnet.org/misq/vol41/iss1/2`

Ramos, I. N., Ramos, K. N., & Ramos, K. S. (2019, 12). Driving the precision medicine highway: community health workers and patient navigators. *Journal of Translational Medicine*, *17*(1), 85. Retrieved from https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-019-1826-2 doi: 10.1186/s12967-019-1826-2

Reimers, N., & Gurevych, I. (2019, 10). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1908.10084

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, *123*(3), 12–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7582737

Ritter, A., Downey, D., Soderland, S., & Etzioni, O. (2008). It's a contradiction—no, it's not: a case study using functional relations. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 11–20).

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., & Lazar, A. J. (2018). Overview of the TREC 2018 Precision Medicine Track. In *Trec.* Gaithersburg, MD.

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Hersh, W. R., Bedrick, S., Lazar, A. J., & Pant, S. (2017). Overview of the TREC 2017 Precision Medicine Track. In *Proceedings of the twenty-sixth text retrieval conference.*

Romanov, A., & Shivade, C. (2018). Lessons from Natural Language Inference in the Clinical Domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1586–1596. Retrieved from https://aclanthology.info/papers/D18-1187/d18-1187

Ross, M. K., Wei, W., & Ohno-Machado, L. (2014, 8). "Big data" and the electronic health record. *Yearbook of medical informatics*, *9*, 97–104. doi: 10.15265/IY-2014-0003

Runeson, P., & Höst, M. (2009, 4). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, *14*(2), 131–164. doi: 10.1007/s10664-008-9102-8

Sánchez, D., & Moreno, A. (2007). Learning medical ontologies from the Web. In *Aime workshop on knowledge management for health care procedures* (pp. 32–45).

Sanderson, M. (1998). Accurate user directed summarization from existing tools. In (pp. 45–51). Association for Computing Machinery (ACM). doi: 10.1145/288627.288640

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, *4*(4), 247–375. doi: 10.1561/1500000009

Saracevic, T. (2007, 11). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. Retrieved from `http://doi.wiley.com/10.1002/asi.20681` doi: 10.1002/asi.20681

Sarafraz, F. (2012a). *Finding conflicting statements in the biomedical literature* (Unpublished doctoral dissertation). University of Manchester.

Sarafraz, F. (2012b). *Finding conflicting statements in the biomedical literature* (Unpublished doctoral dissertation). University of Manchester.

Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H., & Marckmann, G. (2013, 12). What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Medical Ethics*, *14*(1), 55. Retrieved from `https://bmcmedethics.biomedcentral.com/articles/10.1186/1472-6939-14-55` doi: 10.1186/1472-6939-14-55

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D15-1036` doi: 10.18653/v1/D15-1036

Schuurman, N., & Leszczynski, A. (2008, 3). Ontologies for bioinformatics. *Bioinformatics and biology insights*, *2*, 187–200. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/19812775http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2735951`

The Science of Informatics — AMIA. (2019). Retrieved from `https://www.amia.org/about-amia/science-informatics`

Seel, N. M. (2012). Experimental and Quasi-Experimental Designs for Research on Learning. In *Encyclopedia of the sciences of learning* (pp. 1223–1229). Springer US. doi: 10.1007/978-1-4419-1428-6{_}716

Sequiera, R., Baruah, G., Tu, Z., Mohammed, S., Rao, J., Zhang, H., & Lin, J. (2017). Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End destion Answering. *arXiv e-prints*.

Shah, T., Rabhi, F., Ray, P., & Taylor, K. (2014, 1). A Guiding Framework for Ontology Reuse in the Biomedical Domain. In *2014 47th hawaii international conference on system sciences* (pp. 2878–2887). IEEE. doi: 10.1109/HICSS.2014.360

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher,

D. (2007, 8). How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine*, *147*(4), 224–233. doi: 10.7326/0003-4819-147-4-200708210-00179

Simmons, M., Singhal, A., & Lu, Z. (2016, 11). Text mining for precision medicine: Bringing structure to ehrs and biomedical literature to understand genes and health. In *Advances in experimental medicine and biology* (Vol. 939, pp. 139–166). Springer New York LLC. doi: 10.1007/978-981-10-1503-8{_}7

Singhal, A., Leaman, R., Catlett, N., Lemberger, T., McEntyre, J., Polson, S., ... Lu, Z. (2016). Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges. *Database*, *2016*, baw161. Retrieved from http://dx.doi.org/10.1093/database/baw161 doi: 10.1093/database/baw161

Singhal, A., Simmons, M., & Lu, Z. (2016, 7). Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *Journal of the American Medical Informatics Association*, *23*(4), 766–772. Retrieved from https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocw041 doi: 10.1093/jamia/ocw041

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. In *Acm sigir 2008 - 31st annual international acm sigir conference on research and development in information retrieval, proceedings* (pp. 147–154). doi: 10.1145/1390334.1390362

Smucker, M. D., & Jethani, C. P. (2010). Human performance and retrieval precision revisited. In *Sigir 2010 proceedings - 33rd annual international acm sigir conference on research and development in information retrieval* (pp. 595–602). doi: 10.1145/1835449.1835549

Sogancioglu, G., Öztürk, H., & Özgür, A. (2017, 7). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics (Oxford, England)*, *33*(14), i49-i58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28881973 doi: 10.1093/bioinformatics/btx238

Sox, H. C., & Greenfield, S. (2009). Comparative effectiveness research: a report from the institute of medicine. *Annals of internal medicine*, *151*(3), 203–205.

S. Tawfik, N., & R. Spruit, M. (2019). Towards Recognition of Textual Entailment in the Biomedical Domain. In *International conference on applications of natural language to information systems.* Manchester: Springer.

Stelter, R., & Stelter, R. (2014). Professional Practice Between Research, Knowledge and Reflection. In *A guide to third generation coaching* (pp. 175–194). Springer Netherlands. doi: 10.1007/978-94-007-7186-4{_}6

Talman, A., & Chatzikyriakidis, S. (2018). *Testing the Generalization Power of Neural Network Models Across NLI Benchmarks* (Tech. Rep.).

Tamine, L., & Chouquet, C. (2017, 3). On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing and Management*, *53*(2), 332–350. doi: 10.1016/j.ipm.2016.11.004

Tawfik, N., & Spruit, M. (2019, 9). UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain. In *Proceedings of the 18th bionlp workshop and shared task* (pp. 493–499). Association for Computational Linguistics (ACL). doi: 10.18653/v1/w19-5053

Tawfik, N. S., & Spruit, M. R. (2018a). Automated Contradiction Detection in Biomedical Literature. In *International conference on machine learning and data mining in pattern recognition* (Vol. 10934 LNAI, pp. 138–148). Newyork, USA: Springer cham. doi: 10.1007/978-3-319-96136-1{_}12

Tawfik, N. S., & Spruit, M. R. (2018b, 1). The SNPcurator: Literature mining of enriched SNP-disease associations. *Database*, *2018*(2018). doi: 10.1093/database/bay020

Tawfik, N. S., & Spruit, M. R. (2019a). PreMedOnto: A Computer Assisted Ontology for Precision Medicine. In *International conference on applications of natural language to information systems* (Vol. 11608 LNCS, pp. 329–336). Manchester, UK: Springer Verlag.

Tawfik, N. S., & Spruit, M. R. (2019b). Towards Recognition of Textual Entailment in the Biomedical Domain. In *International conference on applications of natural language to information systems* (Vol. 11608 LNCS, pp. 368–375). Manchester, UK: Springer Verlag. doi: 10.1007/978-3-030-23281-8{_}32

Tawfik, N. S., & Spruit, M. R. (2020a). Computer-assisted Relevance Assessment: a Case-study of Updating Systematic Reviews. *Applied Sciences*.

Tawfik, N. S., & Spruit, M. R. (2020b). Evaluating Sentence Representations for Biomedical Text: Methods and Experimental Results. *Journal of Biomedical Informatics*.

Tedre, M., & Moisseinen, N. (2014). Experiments in computing: a survey. *The Scientific World Journal*, *2014*, 1–11. doi: 10.1155/2014/549398

Thermes, C. (2014, 9). Ten years of next-generation sequencing technology. *Trends in genetics : TIG*, *30*(9), 418–426. doi: 10.1016/j.tig.2014.07.001

Thomas, G. (2011, 7). A Typology for the Case Study in Social Science Following a Review of Definition, Discourse, and Structure. *Qualitative In-*

quiry, *17*(6), 511–521. Retrieved from http://journals.sagepub.com/doi/10.1177/1077800411409884 doi: 10.1177/1077800411409884

Thomas, P., Rocktäschel, T., Hakenberg, J., Lichtblau, Y., & Leser, U. (2016, 6). {SETH} detects and normalizes genetic variants in text. *Bioinformatics*, btw234. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btw234 doi: 10.1093/bioinformatics/btw234

Thomas, P. E., Klinger, R., Furlong, L. I., Hofmann-Apitius, M., & Friedrich, C. M. (2011). Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers. *BMC bioinformatics*, *12*(4), 1.

Tombros, A., Sanderson, M., & Gray, P. (1998). Advantages of query biased summaries in information retrieval. In *Sigir* (Vol. 98, pp. 2–10).

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., ... Paliouras, G. (2015, 12). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, *16*(1), 138. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6 doi: 10.1186/s12859-015-0564-6

Vamathevan, J., & Birney, E. (2017, 9). A Review of Recent Advances in Translational Bioinformatics: Bridges from Biology to Medicine. *Yearbook of Medical Informatics*, *26*(01), 178–187. Retrieved from http://www.thieme-connect.de/DOI/DOI?10.15265/IY-2017-017 doi: 10.15265/IY-2017-017

Van Der Vegt, A., Zuccon, G., Koopman, B., & Deacon, A. (2019, 5). Impact of a search engine on clinical decisions under time and system effectiveness constraints: Research protocol. *Journal of Medical Internet Research*, *21*(5). doi: 10.2196/12803

Vickery, B. C. (1959a). The structure of information retrieval systems. In *Proceedings of the international conference on scientific information* (Vol. 2, pp. 1275–1290).

Vickery, B. C. (1959b). Subject analysis for information retrieval. In *Proceedings of the international conference on scientific information* (Vol. 2, pp. 855–865).

Wang, J. (2011). Accuracy, agreement, speed, and perceived difficulty of users' relevance judgments for e-discovery. In *Proceedings of sigir information retrieval for e-discovery workshop* (Vol. 1).

Wang, J., & Soergel, D. (2010). A user study of relevance judgments for e-discovery. In *Proceedings of the 73rd asis&t annual meeting on navigating streams in an information ecosystem-volume 47* (p. 74).

Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., &

Liu, H. (2018, 10). MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 1–16. Retrieved from `http://link.springer.com/10.1007/s10579-018-9431-1` doi: 10.1007/s10579-018-9431-1

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., ... Liu, H. (2018, 11). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, *87*, 12–20. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1532046418301825?via%3Dihub` doi: 10.1016/J.JBI.2018.09.008

Wang, Y., Liu, S., Rastegar-Mojarad, M., Afzal, N., Wang, L., Shen, F., ... Liu, H. (2018). Overview of BioCreative/OHNLP Challenge 2018 Task 2: Clinical Semantic Textual Similarity. In *Proceedings of the biocreative/ohnlp challeng.* Washington. Retrieved from `https://github.com/ohnlp/BioCreativeOHNLPProceedings/raw/master/clinicalsts_overview.pdf` doi: 10.13140/RG.2.2.26682.24006

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... Liu, H. (2018, 1). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. doi: 10.1016/j.jbi.2017.11.011

Wei, C.-H., Harris, B. R., Kao, H.-Y., & Lu, Z. (2013, 6). tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, *29*(11), 1433–1439. Retrieved from `http://dx.doi.org/10.1093/bioinformatics/btt156` doi: 10.1093/bioinformatics/btt156

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., ... others (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, *42*(D1), D1001–D1006.

Weng, W.-H., & Szolovits, P. (2019, 9). Representation Learning for Electronic Health Records. Retrieved from `http://arxiv.org/abs/1909.09248`

White, L., Togneri, R., Liu, W., Bennamoun, M., & Ben, M. (2015). How Well Sentence Embeddings Capture Meaning. In *Proceedings of the 20th australasian document computing.* Parramatta, NSW, Australia: ACM. Retrieved from `http://dx.doi.org/10.1145/2838931.2838932` doi: 10.1145/2838931.2838932

Wiederhold, G., & McCarthy, J. (2010, 4). Arthur Samuel: Pioneer in Machine Learning. *IBM Journal of Research and Development*, *36*(3), 329–331. doi: 10.1147/rd.363.0329

Wieting, J., & Kiela, D. (2019). No Training Required: Exploring Random Encoders for Sentence Classification. In *International conference* *on learning representations.* Retrieved from https://openreview.net/forum?id=BkgPajAcY7

Williams, A., Nangia, N., & Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (p. 1112–1122). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-1101

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2000). *Experimentation in Software Engineering* (Vol. 6). Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-1-4615-4625-2 doi: 10.1007/978-1-4615-4625-2

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, *1*(1), 67. Retrieved from https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf

Wu, S., & Liu, H. (2011). Semantic characteristics of NLP-extracted concepts in clinical notes vs. biomedical literature. In *Amia ... annual symposium proceedings / amia symposium. amia symposium* (Vol. 2011, pp. 1550–1558).

Wynn, R. M., Adams, K. T., Kowalski, R. L., Shivega, W. G., Ratwani, R. M., & Miller, K. E. (2018). The Patient in Precision Medicine: A Systematic Review Examining Evaluations of Patient-Facing Materials. *Journal of Healthcare Engineering*, *2018*. doi: 10.1155/2018/9541621

Xiang, Z., Courtot, M., Brinkman, R. R., Ruttenberg, A., & He, Y. (2010, 6). OntoFox: web-based support for ontology reuse. *BMC Research Notes*, *3*(1), 175. doi: 10.1186/1756-0500-3-175

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., & Xu, H. (2015). Citation Sentiment Analysis in Clinical Trial Papers. *AMIA Annual Symposium proceedings. AMIA Symposium*, *2015*, 1334–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26958274

Yaghoobzadeh, Y., & Schütze, H. (2016). Intrinsic Subspace Evaluation of Word Embedding Representations. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 236–246). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P16-1023 doi: 10.18653/v1/P16-1023

Yang, L., Ai, Q., Spina, D., Chen, R.-C., Pang, L., Croft, W. B., ... Scholer, F. (2016). Beyond factoid qa: Effective methods for non-factoid answer sentence retrieval. In *European conference on information retrieval* (pp.

115–128).

Yang, Z., Zhu, C., & Chen, W. (2019, 11). Parameter-free Sentence Embedding via Orthogonal Basis. In (pp. 638–648). Association for Computational Linguistics (ACL). doi: 10.18653/v1/d19-1059

Yates, L. R., Seoane, J., Le Tourneau, C., Siu, L. L., Marais, R., Michiels, S., ... Andre, F. (2018, 1). The European Society for Medical Oncology (ESMO) Precision Medicine Glossary. *Annals of Oncology*, *29*(1), 30–35. doi: 10.1093/annonc/mdx707

Yepes, A. J., & Verspoor, K. (2014, 6). Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. *F1000Research*. Retrieved from `http://dx.doi.org/10.12688/f1000research.3-18.v2` doi: 10.12688/f1000research.3-18.v2

Yin, R. k. (2017). *Case Study Research and Applications: Design and Methods, Sixth Edition.*

Zadrozny, W., & Garbayo, L. (2018). A Sheaf Model of Contradictions and Disagreements. Preliminary Report and Discussion. In *International symposium on artificial intelligence and mathematics,*. Florida.

Zelkowitz, M. V., & Wallace, D. R. (1998, 5). Experimental models for validating technology. *Computer*, *31*(5), 23–31. doi: 10.1109/2.675630

Zhang, H., Abualsaud, M., Ghelani, N., Smucker, M. D., Cormack, G. V., & Grossman, M. R. (2018, 10). Effective user interaction for high-recall retrieval: Less is more. In *International conference on information and knowledge management, proceedings* (pp. 187–196). Association for Computing Machinery. doi: 10.1145/3269206.3271796

Zhang, H., Cormack, G. V., Grossman, M. R., & Smucker, M. D. (2019). Evaluating sentence-level relevance feedback for high-recall information retrieval. *Information Retrieval Journal*. doi: 10.1007/s10791-019-09361-0

List of publications

1. Tawfik, N. S., & Spruit, M. R. (2020a) Computer-assisted Relevance Assessment: a Case-study of Updating Systematic Reviews. *Applied Sciences*, *10*(8), 2845.
2. Tawfik, N. S., & Spruit, M. R. (2020b) Evaluating Sentence Representations for Biomedical Text: Methods and Experimental Results. *Journal of Biomedical Informatics*, 103396.
3. Tawfik, N. S., & Spruit, M. R. (2019) UU_TAILS at MEDIQA 2019: Learning Textual Entailment in the Medical Domain. In *Proceedings of the 18th BioNLP workshop and Shared Task* (pp. 493–499).
4. Tawfik, N. S., & Spruit, M. R. (2019a). PreMedOnto: A Computer Assisted Ontology for Precision Medicine. In *International Conference on Applications of Natural Language to Information Systems* (pp. 329–336). Springer, Cham.
5. Tawfik, N. S., & Spruit, M. R. (2019b) Towards Recognition of Textual Entailment in the Biomedical Domain. In *International Conference on Applications of Natural Language to Information Systems* (pp. 368–375). Springer, Cham.
6. Tawfik, N. S., & Spruit, M. R. (2018a). Automated Contradiction Detection in Biomedical Literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 138–148). Springer, Cham.
7. Tawfik, N. S., & Spruit, M. R. (2018b) The SNPcurator: Literature mining of enriched SNP-disease associations. *Database, 2018.*
8. Tawfik, N. S., Youssef, S. M., Kholief, M. (2016). A hybrid automated detection of epileptic seizures in EEG records. *Computers Electrical Engineering, 53*, 177-190.

Summary

In recent years, the precision medicine (PM) approach has emerged to mitigate the shortcomings of the traditional ”one-size-fits-all” that has dominated the medical practice for so long. Precision medicine aims at developing more precise and tailored plans for patients starting from screening and diagnosing to treatment and interventions. This goes beyond the classical ”signs-and-symptoms” method by taking into consideration patients’ genetics, biomarkers, lifestyle, and further environmental influences. With the advances in genetics and the availability of health data, practitioners and scientists are now able to transform the precision medicine paradigm into a clinical reality. Realistically, achieving this requires the integration of different sources of data such as scientific literature, electronic health records, and structured databases through the means of big data analyticS.

This thesis investigates the role of Natural Language Processing (NLP) or, more specifically, Biomedical NLP (BioNLP), in the precision medicine revolution. The potential benefit of extracting valuable information from the existing unstructured data improves the understanding of the field, leading to better healthcare services and an overall increase in patients’ satisfaction. The main research question of this Ph.D. thesis, therefore, is:
How can Biomedical NLP techniques support and advance the precision medicine approach through collection and analysis of clinical and medical textual resources?
This research is a step forward towards the application of Precision Medicine by focusing on two main goals: advancing the PM applicability and proving the efficacy of the paradigm discussed in part one and two of this dissertation, respectively. In total, this work resulted in seven papers, mapped to the information systems research framework. The first part of the dissertation (chapters 2 and 3) provides tools to address the needs of the research experts, whereas the second part (chapters 4-8) investigates transfer learning models to fit the required BioNLP tasks.

Since precision medicine was originally empowered by the completion of the Human Genome Project, the first contribution aimed at automating the curation process of published genetic literature, chapter 2 provides a description

of the SNPcurator online tool that allows users to browse Genome-Wide Association Studies. It enables users to query specific diseases and view related single nucleotide polymorphisms (SNPs) according to cohort and statistical significance. The tool relies on a fully automated NLP framework and is up-to-date with the daily content added to the PubMed repository. Chapter 3 presents PreMedOnto, an ontology designed to organize the hierarchy of the PM domain and its general investigations, diagnostics, and treatment terminologies. The ontology development followed a reuse-based approach to map terms and concepts from gold standard biomedical ontologies.

The second part of this research focuses on analyzing clinical interventions by combining NLP with machine learning methods. Chapter 4 extracts published findings, concerning a medical intervention in the cardiovascular disease, from abstracts and highlights conflicts in the literature about the efficacy of the practice. This work is extended in Chapter 5, by not only extracting contradictions but also by detecting Natural Language Inferences (NLIs) in the biomedical domain. NLI is a core NLP task that models the relation between two input texts, i.e., given two snippets of text, Premise P and Hypothesis H, textual entailment recognition determines if the meaning of H can be inferred from that of P. As a first attempt to model inference in the biomedical domain, we enrich the previously used dataset to align with the standard format of NLI benchmarks. We explore traditional features and dedicated-sentence encoders and evaluate their performances. Moreover, we scale-up the evaluation to investigate the gain of combining hand-crafted features and sentence representations in a deep neural network. However, a major limitation to the development of robust NLP models in the biomedical domain is the annotation bottleneck that exists widely in the biomedical field. The previous chapters proposed models that were trained, tested, validated, and reported but on a rather small dataset. Fortunately, in late 2019 the MedNLI benchmark was published as part of the MEDIQA challenge organized by the BioNLP workshop. Prior to participating in the challenge, we performed an exploratory embedding analysis, as depicted in Chapter 6, to investigate the ability of existing embedding methods in capturing the semantics of clinical sentences. We explored different methods (static, contextualized, dedicated sentence encoders) across various medical NLP tasks in different datasets. Chapter 7 describes our participation in the NLI and RQE tasks of the MEDIQA challenge to detect inference between pairs of sentences and questions, respectively.

Finally, Chapter 8 is a real-world case study on how NLI may be employed to assist medical experts in high-recall retrieval tasks. We conduct a controlled experiment to assist with the systematic review update process. The case study aims at speeding-up the assessment process of medical experts in judging the relevancy of scientific articles while maintaining a good recall.

Samenvatting

In de afgelopen jaren is de benadering van precisiegeneeskunde (‘Precision Medicine’; PM) tot bloei gekomen als mogelijke oplossing voor de tekortkomingen van de traditionele ‘one-size-fits-all’ benadering die de medische praktijk zo lang heeft gedomineerd. Precisiegeneeskunde is gericht op het ontwikkelen van nauwkeurigere en op maat gemaakte behandelplannen voor patiënten, van screening en diagnose tot behandeling en interventies. Dit gaat verder dan de klassieke ”tekenen-en-symptomen-methode door rekening te houden met de genetica, biomarkers, levensstijl en andere omgevingsinvloeden van patiënten. Met de vooruitgang in de genetica en de beschikbaarheid van gezondheidsgegevens zijn artsen en wetenschappers nu in staat om het paradigma van precisiegeneeskunde om te zetten in een klinische realiteit. Praktisch gezien vereist dit de integratie van verschillende gegevensbronnen, zoals wetenschappelijke literatuur, elektronische patiëntendossiers en gestructureerde databases door middel van grootschalige data-analyse (‘Big Data Analytics’). Dit proefschrift onderzoekt de rol van natuurlijke taalverwerking (‘Natural Language Processing’; NLP), of meer specifiek biomedische NLP (BioNLP), in de context van de revolutie rondom precisiegeneeskunde. Het potentiële voordeel van het extraheren van waardevolle informatie uit bestaande ongestructureerde gegevens verbetert het begrip van het veld, wat leidt tot betere gezondheidszorg en algehele verbetering van de tevredenheid van de patiënt. De hoofdonderzoeksvraag van dit proefschrift is daarom: **Hoe kunnen biomedische NLP-technieken de benadering van precisiegeneeskunde ondersteunen en bevorderen door het verzamelen en analyseren van klinische en medische tekstuele bronnen?**

Dit onderzoek is een stap voorwaarts in de toepassing van precisiegeneeskunde door de focus op twee hoofddoelen: het bevorderen van de PM-toepasbaarheid en het aantonen van de doeltreffendheid van het paradigma, besproken in respectievelijk deel één en twee van dit proefschrift. In totaal heeft dit werk geresulteerd in zeven publicaties, die gepositioneerd zijn in het onderzoekskader van informatiesystemen. Het eerste deel van het proefschrift (hoofdstukken 2 en 3) biedt hulpmiddelen om aan de behoeften van de onderzoeksexperts te voldoen, terwijl het tweede deel (hoofdstukken 4-8) modellen voor de overdracht van het automatisch geleerde (‘Transfer Learning’) onderzoekt die passen bij de vereiste BioNLP-taken. Aangezien precisiegeneeskunde

oorspronkelijk werd bekrachtigd door de voltooiing van het Human Genome Project, is de eerste bijdrage gericht op het automatiseren van het curatieproces van gepubliceerde genetische literatuur. Hoofdstuk 2 geeft een beschrijving van de SNPcurator webtoepassing waarmee gebruikers kunnen bladeren door genoom-brede associatiestudies ('Genome-Wide Association Studies'; GWAS). Het stelt gebruikers in staat om specifieke ziekten te bevragen en gerelateerde enkel-nucleotide polymorfieën ('Single Nucleotide Polymorfism'; SNP) te bekijken op basis van cohort en statistische significantie. De SNPcurator is gebaseerd op een volledig geautomatiseerd NLP-raamwerk en is altijd bijgewerkt met de dagelijkse uitbreidingen en wijzigingen die aan de PubMed-bibliotheek worden toegevoegd. Hoofdstuk 3 presenteert PreMedOnto, een ontologie die is ontworpen om de hiërarchie van het precisiegeneeskunde domein en de algemene onderzoeks-, diagnostiek- en behandelingsterminologieën te organiseren. De ontologie-ontwikkeling volgde een op hergebruik gebaseerde benadering om termen en concepten uit gouden standaard biomedische ontologieën in kaart te brengen.

Het tweede deel van deze dissertatie richt zich op het analyseren van klinische interventies door NLP te combineren met methoden voor automatisch leren ('Machine Learning'; ML). Hoofdstuk 4 haalt gepubliceerde bevindingen over een medische interventie bij cardiovasculaire aandoeningen uit gepubliceerde abstracts en belicht conflicten in de literatuur over de effectiviteit van de praktijk. Dit werk wordt uitgebreid in Hoofdstuk 5, door niet alleen tegenstrijdigheden te extraheren, maar ook door inferenties in natuurlijke taal ('Natural Language Inference'; NLI) in het biomedische domein te detecteren. NLI is een kerntaak van NLP die de relatie tussen twee invoerteksten modelleert, dat wil zeggen gegeven twee tekstfragmenten, premisse P en hypothese H, wil men bepalen of de betekenis van H kan worden afgeleid uit die van P ('Textual Entailment'). Als een eerste poging om inferentie te modelleren in het biomedische domein verrijken we de in Hoofdstuk 4 reeds beschreven dataset om deze af te stemmen op het standaardformaat van NLI vergelijkende studies ('NLI Benchmark'). We verkennen zowel traditionele functies als toegewijde zinscoderingen ('Sentence Encoder') en evalueren hun prestaties. Bovendien schalen we de evaluatie op om de meerwaarde te onderzoeken van het combineren van handgemaakte functies en zinsrepresentaties in een diep neuraal netwerk. Een belangrijke beperking voor de ontwikkeling van robuuste NLP-modellen in het biomedische domein is echter het knelpunt van het slechts beperkt op grote schaal beschikbaar zijn van geannoteerde gegevensbanken in het biomedische veld. In de vorige hoofdstukken werden namelijk modellen voorgesteld die waren getraind, getest, gevalideerd en gerapporteerd op een relatief kleine dataset. Gelukkig werd eind 2019 de MedNLI-benchmark gepubliceerd als onderdeel van de MEDIQA-uitdaging binnen de wereldwijde BioNLP-werkgroep. Voordat we aan deze uitdaging deelnamen, hebben we een verkennende analyse van verscheidene inbeddingsmethoden uitgevoerd, zoals beschreven in Hoofdstuk 6, om het vermogen van bestaande inbeddingsmethoden voor het opnemen van de inhoudelijke betekenis van klinische zinnen te onderzoeken. We hebben verschillende methoden onderzocht (statische, gecontextualiseerde, en toegewijde zinscoderingen) voor verschillende medische NLP-taken in verschillende datasets. Hoofdstuk 7 beschrijft onze deelname aan de NLI- en de vraag-gelijkenis- ('Recognizing Question Entailment'; RQE)-taken van de MEDIQA-uitdaging om gevolgtrekkingen te detecteren tussen respectievelijk paren zinnen en vragen. Ten slotte is Hoofdstuk 8 een praktijkgerichte casus rondom het gebruik van NLI om medische experts te helpen bij het zoeken naar informatie waarbij het niet missen van relevante stukken centraal staat ('High-Recall'). We voeren een gecontroleerd experiment uit om te helpen bij het bijwerkingsproces voor systematische literatuurbeoordelingen. De casus heeft tot doel het beoordelingsproces voor medische experts te versnellen bij het beoordelen van de relevantie van wetenschappelijke artikelen, teneinde systematische literatuurbeoordelingen efficiënter en effectiever periodiek bij te kunnen werken, wat speciale aandacht vereist voor het niet abusievelijk missen van relevante stukken.

Curriculum Vitae

Noha Seddik Tawfik was born on November 5th, 1988, in Alexandria, Egypt. She attended the Arab Academy For Science, Technology and Maritime Transport where she obtained a Bachelor's and a Master's degrees in Computer Engineering in 2011 and 2014, respectively. She wrote her Master's thesis on the *Automated Detection of Epileptic Seizure in EEG Integrating Weighted Permutation Entropy and Support Vector Machine Classification*, a biomedical problem that required analysis of patient's EEG data to determine the onset zone defined as the focal starting point of epileptic seizure.

In 2016, she started her research as a part-time PhD candidate at the department of Information and Computing Sciences of Utrecht University. Noha worked remotely from Egypt and scheduled regular video calls with her supervisor to discuss research progress. She also visited Utrecht University twice per year for a period of a month each time. During her visits, Noha joins the department's research meetings and the ADS lab colloquia. In the course of her time as a PhD researcher, She presented her work at various international scientific meetings, including the BioNLP workshop co-located with the 62nd Annual Meeting of the Association of Computational Linguistics. Simultaneous to doing her research, Noha has been acting as a teaching assistant at the department of Computer Engineering in her home university. Among her teaching duties were preparing and conducting tutorial and lab sessions for various bachelor courses including Structured Programming, Systems Programming and Pattern Recognition.