TUCS

Juho Heimonen

Knowledge Representation and Text Mining in Biomedical, Healthcare, and Political Domains

Turku Centre for Computer Science

TUCS Dissertations
No 246, November 2019

Knowledge Representation and Text Mining in Biomedical, Healthcare, and Political Domains

Juho Heimonen

To be presented, with the permission of the Faculty of Science and Engineering of the University of Turku, for public criticism in the lecture hall XX on November 27, 2019, at 12 noon.

University of Turku
Department of Future Technologies
FI-20014 Turun yliopisto
Finland

2019

Supervisors

Prof. Tapio Salakoski, PhD
Department of Future Technologies
University of Turku
Turku, Finland

Prof. Sanna Salanterä, PhD, RN
Department of Nursing Science
University of Turku
Turku, Finland

Assoc. Prof. Tapio Pahikkala, PhD
Department of Future Technologies
University of Turku
Turku, Finland

Reviewers

Prof. Martti Juhola, PhD
Computing Sciences Unit
Tampere University
Tampere, Finland

Lec. Sumithra Velupillai, PhD
Department of Psychological Medicine
King's College London
London, United Kingdom

Opponent

Assoc. Prof. Øystein Nytrø, PhD
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-952-12-3881-9
ISSN 1239-1883

Abstract

Knowledge representation and text mining can be employed to discover new knowledge and develop services by using the massive amounts of text gathered by modern information systems. The applied methods should take into account the domain-specific nature of knowledge. This thesis explores knowledge representation and text mining in three application domains.

Biomolecular events can be described very precisely and concisely with appropriate representation schemes. Protein–protein interactions are commonly modelled in biological databases as binary relationships, whereas the complex relationships used in text mining are rich in information. The experimental results of this thesis show that complex relationships can be reduced to binary relationships and that it is possible to reconstruct complex relationships from mixtures of linguistically similar relationships. This encourages the extraction of complex relationships from the scientific literature even if binary relationships are required by the application at hand. The experimental results on cross-validation schemes for pair-input data help to understand how existing knowledge regarding dependent instances (such those concerning protein–protein pairs) can be leveraged to improve the generalisation performance estimates of learned models.

Healthcare documents and news articles contain knowledge that is more difficult to model than biomolecular events and tend to have larger vocabularies than biomedical scientific articles. This thesis describes an ontology that models patient education documents and their content in order to improve the availability and quality of such documents. The experimental results of this thesis also show that the Recall-Oriented Understudy for Gisting Evaluation measures are a viable option for the automatic evaluation of textual patient record summarisation methods and that the area under the receiver operating characteristic curve can be used in a large-scale sentiment analysis. The sentiment analysis of Reuters news corpora suggests that the Western mainstream media portrays China negatively in politics-related articles but not in general, which provides new evidence to consider in the debate over the image of China in the Western media.

Tiivistelmä

Tiedonesitystapojen ja tekstinlouhinnan avulla nykyajan tietojärjestelmien keräämiä valtavia tekstivarantoja voidaan käyttää uuden tiedon tuottamiseen ja palveluiden kehittämiseen. Käytettävien menetelmien on kuitenkin huomioitava tiedon alakohtainen luonne. Tässä väitöskirjassa tarkastellaan tiedonesitystapoja ja tekstinlouhintaa kolmella eri sovellusalalla.

Sopivilla tiedonesitystavoilla biomolekyylien välisiä tapahtumia voidaan kuvata hyvin täsmällisesti ja tiiviisti. Proteiinien välisiä vuorovaikutuksia kuvataan biotieteellisissä tietokannoissa yleensä kahdenvälisinä vuorovaikutuksina, kun taas tekstinlouhinnassa käytettävät monenväliset vuorovaikutukset sisältävät runsaasti informaatiota. Väitöskirjan kokeelliset tulokset osoittavat, että monenvälisistä vuorovaikutuksista on mahdollista johtaa kahdenvälisiä vuorovaikutuksia ja että kielellisesti samanlaisten vuorovaikutusten sekoituksista voidaan erottaa yksittäiset monenväliset vuorovaikutukset. Nämä havainnot puoltavat monenvälisten vuorovaikutusten uuttamista tieteellisestä kirjallisuudesta. Lisäksi parittaisella datalla tehdyt ristiinvalidointikokeet auttavat ymmärtämään, miten datassa olevat riippuvuudet voidaan ottaa tunnetun tiedon avulla paremmin huomioon opittujen mallien suorituskykyä arvioitaessa.

Uutisartikkeleissa ja terveydenhuollon tuottamissa asiakirjoissa käytetään kieltä monipuolisemmin ja käsitellään laajempia aiheita kuin biolääketieteellisissä artikkeleissa, ja osittain juuri siksi niissä olevaa tietoa on vaikeampi mallintaa kuin biomolekyylien välisiä tapahtumia. Väitöskirja esittelee potilasohjauksessa käytettävien kirjallisten ohjeiden saatavuuden ja laadun parantamiseen tarkoitetun ontologian, jolla kuvataan ohjeita ja niiden sisältöä. Kokeelliset tulokset puolestaan osoittavat, että ROUGE-mitat soveltuvat potilasasiakirjojen tiivistämismenetelmien automaattiseen arviointiin ja että ROC-käyrän alle jäävää pinta-alaa voidaan käyttää mielipiteiden louhinnassa suuresta aineistosta. Mielipiteiden louhinnalla uutisia sisältävästä tekstikokoelmasta saadut tulokset viittavat siihen, että länsimaisella valtavirtauutismedialla on taipumus kuvata Kiinaa verrokkimaita negatiivisemmin politiikka-aiheisissa uutisissa mutta ei koko uutisvirrassa. Nämä havainnot tuovat oman näkökulmansa keskusteluun Kiinan imagosta länsimaisessa mediassa.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors Tapio Salakoski, Sanna Salanterä, and Tapio Pahikkala who guided me through my doctoral studies. I am particularly grateful for their invaluable support, encouragement, and patience without which the completion of this thesis would not have been possible. I have enjoyed our many discussions ranging from practical matters to theoretical and ethical topics. I am also grateful to the reviewers of this thesis, Martti Juhola and Sumithra Velupillai, for their constructive criticisms and helpful comments with which I was able to improve the thesis and to Øystein Nytrø for agreeing to act as my opponent.

I would like to thank Filip Ginter and Sampo Pyysalo for guiding me as I dived into the world of research and Antti Airola, Jukka Aukia, Jari Björne, Riitta Danielsson-Ojala, Heljä Lundgrén-Laine, Hans Moen, and Laura-Maria Peltonen for the collaboration that was crucial for the completion of this thesis. I also want to extend my thanks to Suwisa Kaewphan, Farrokh Mehryary, and Timo Viljanen for the supportive and lively discussions over the years, and to all my other colleagues in the IKITIK consortium and at the Department of Future Technologies with whom I have had the pleasure to work and who are too numerous to mention here.

I am grateful to TUCS - Turku Centre for Computer Science for the funding and infrastructure which have greatly supported my studies. I acknowledge the Academy of Finland and the Finnish Funding Agency for Innovation for financial support and the CSC - IT Center for Science for computational resources. I also wish to thank the IKITIK consortium for the motivating environment and Lingsoft Oy for the long-term collaboration.

Last, I want to express my warmest thanks to my wife Oili, whose patience and love has supported me during my studies, and to my two lovely sons, Matias and Aatos, who have brought much joy to my life.

Turku, October 2019
Juho Heimonen

List of original publications

I Heimonen, J., Pyysalo, S., Ginter, F. and Salakoski, T. (2008). Complex-to-pairwise mapping of biological relationships using a semantic network representation. In Salakoski, T., Rebholz-Schuhmann, D. and Pyysalo, S. (eds.), *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (pp. 45–52). Turku, Finland: Turku Centre for Computer Science

II Heimonen, J., Björne, J. and Salakoski, T. (2010). Reconstruction of semantic relationships from their projections in biomolecular domain. In Cohen, K. B., Demner-Fushman, D., Ananiadou, S., Pestian, J., Tsujii, J. and Webber, B. (eds.), *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (pp. 108–116). Stroudsburg, PA, USA: Association for Computational Linguistics

III Heimonen, J., Salakoski, T. and Pahikkala, T. (2014). Properties of object-level cross-validation schemes for symmetric pair-input data. In Fränti, P., Brown, G., Loog, M., Escolano, F. and Pelillo, M. (eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 384–393). Berlin Heidelberg, Germany: Springer

IV Heimonen, J., Danielsson-Ojala, R., Salakoski, T., Lundgrén-Laine, H. and Salanterä, S. (2018). Ontology development for patient education documents using a professional- and patient-oriented Delphi method. *CIN: Computers, Informatics, Nursing* 36:448–457

V Moen, H., Peltonen, L.-M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T. and Salanterä, S. (2016). Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine* 67:25–37

VI Aukia, J., Heimonen, J., Pahikkala, T. and Salakoski, T. (2017). Automated quantification of Reuters news using a receiver operating characteristic curve analysis: the Western media image of China. *Global Media and China* 2:251–268

Reprints of the original publications are only available in the printed version.

Contents

I Research summary 1

1 Introduction 3
1.1 Research questions and contributions 8
1.2 Structure of thesis 9

2 Domain-specific nature of text and knowledge 11
2.1 Biomedical domain 12
2.2 Healthcare domain 18
2.3 Political domain 21

3 Thesis contributions 25
3.1 Complex-to-binary mapping of protein–protein relationships . 25
3.2 Reconstruction of protein–protein relationships 28
3.3 Cross-validation on pair-input data 31
3.4 Ontology for patient education documents 35
3.5 Summarisation of electronic patient records 38
3.6 Large-scale sentiment analysis of news articles 40

4 Summary 43

II Original publications 63

Part I

Research summary

Chapter 1

Introduction

Modern information systems store and process enormous quantities of data in various forms. Much of these data is primarily produced for documentation, archival, or information exchange purposes. This applies particularly to data in the form of natural language. Newspaper and scientific articles are published to share information, patient records are written to document patients' medical care, and scientific discoveries are archived into specialised databases for further analysis. The large scale of these data offers valuable opportunities to discover new knowledge that is not available in a structured form or directly evident from individual pieces of data.

This thesis focuses on two aspects of refining textual data into human knowledge: knowledge representation and text mining. The former provides means to express human knowledge in a computer-readable manner, whereas the latter provides tools to obtain such knowledge from textual data. They are discussed in the context of biomedical, healthcare, and political domains. Each of these domains routinely use textual data but the essences of the domains, and hence of the texts as well, are remarkably distinct.

The purpose of knowledge representation is to capture human knowledge of the world, which is typically achieved by modelling the world and expressing the knowledge as statements of facts. The human knowledge is neither perfect nor complete, however, which must be taken into account in knowledge representations and their applications. In the acquisition and analysis of knowledge, one must consider the possibility of experimental errors and misinterpretations of observations, the imprecise nature of human communication, subjectivity of matters, and intentional or unintentional bias to promote particular agenda or opinion. The sources, magnitudes, and consequences of these issues depend on the characteristics of the domain, particularly its subject matter and the objectives of the discourse. They are also reflected in the characteristics of textual data sources and in the methodological choices as seen in the later chapters.

A broad and practical definition of knowledge representation is adopted: a knowledge representation is a construct that expresses human knowledge in such a manner that it is both human-understandable and computer-readable and has utility in practical applications or in research tasks. The technical details of the representation are of secondary importance as long as the representation can be used computationally. The purpose of this demarcation is to direct the focus to the various means by which knowledge can be expressed in information systems in a human-friendly manner and to avoid restricting the discussion to any specific theoretical approach.

The most obvious knowledge representations are symbolic representations that are explored in the field of knowledge representation and reasoning. The field, historically categorised under artificial intelligence, studies how knowledge can be expressed formally with symbolic methods and how mathematical logic can be applied to reason over such knowledge (Brachman and Levesque, 2004). It is closely associated with mathematics, which is the basis for formal methods, and to computer science, which offers tools to implement efficient algorithms. The challenge of symbolic knowledge representations is to create mathematically rigorous models of knowledge for domains in which human knowledge is inherently uncertain or imprecise. Indeed, many recent advances in artificial intelligence have been made with statistical methods developed for high-dimensional data. Nevertheless, symbolic representations remain useful for domain-specific applications in which knowledge can be expressed as sets of rules or statements.

Formal knowledge representations, such as description logics, are tools for constructing ontologies, which explicitly and formally model specific fragments of the real world. Top-level ontologies concern general knowledge and describe concepts that can be used across domains, such as objects, time and processes, whereas specific ontologies model particular domains, tasks, or applications. The common components of ontologies include individual objects and object collections as well as features of and relations between the objects and/or collections. These elements can be used to describe each other, which results in a rich network of connections. (Jakus et al., 2013).

Symbolic knowledge representations are not the only constructs that fit the adopted definition of knowledge representations. Relaxing the formality requirement, there are computer-readable alternatives to symbolic representations which are equally informative for humans but cannot be used for automated formal reasoning. Vocabularies and classifications, for example, are constructs that can be implemented in information systems to effectively manipulate and store human knowledge but the stored information is only fully meaningful to the human users of the systems (Chen et al., 2005). These constructs emphasise the information technological aspect of storing structured information into databases and analysing it manually or algorithmically. This perspective to knowledge is close to that adopted by the field of data mining, which studies the discovery of patterns in large data sets through statistical and machine learning methods.

Given the importance of statistical learning methods in artificial intelligence (Flasiński, 2016), one could argue that statistically induced models, such as word-space models or artificial neural networks, implicitly contain real-world knowledge that humans cannot decode but that can be applied computationally. Such a model does not necessarily align itself with what humans perceive as knowledge but new knowledge may be discovered through the use and analysis of the model. If the model can be utilised by an algorithm to make human-understandable inferences or predictions, should it be classified as computer knowledge (as opposed to human knowledge)? The question whether these models are knowledge representations is beyond the scope of the thesis which, as stated above, has the perspective of utility to humans. Statistical models are considered to be tools to analyse data and extract knowledge but not knowledge representations themselves.

The wealth of textual data stored in information systems may be utilised with natural language processing (NLP) and text mining methods. Computational linguistics is an interdisciplinary field that concerns the computational modelling, analysis, and generation of natural language, both written and spoken. It relies on linguistics, which provides insights into the nature of human language, and computer science, which is the basis of computational analysis methods. NLP is the branch of computational linguistics that focuses on the engineering aspects of natural language. The methods employed in NLP often involve statistical approaches which is why also statistics and machine learning are essential in NLP. (Clark et al., 2013). Text mining, or textual data mining, studies the extraction of structured information from text. It overlaps with NLP as both explore textual data and employ statistical and computational methods. In contrast to NLP, which studies a wide variety of tasks involving natural language, text mining focuses on the data mining and knowledge discovery aspects. (Aggarwal and Zhai, 2012).

Due to the diversity of the aspects of natural language, NLP and text mining methods commonly address one task, or at most a few, that can be considered independently. Basic NLP tasks include the splitting of text strings into sentences and further into tokens followed by the analysis of morphology (i.e. word structure), syntax (i.e. sentence structure), and semantics (i.e. meaning). Text mining builds on the outputs of these analyses (such as tokenisation, lemmatisation, part-of-speech analysis, and syntactic parse) and employs data mining methods to produce structured information. (Ananiadou and McNaught, 2006).

Common text mining tasks include document classification, term or concept extraction and normalisation, named entity recognition, relation and event extraction, text summarisation, and sentiment analysis. In document classification, one or more predefined labels (such as topics or authors) are assigned to text documents. In term/concept extraction and normalisation, pertinent terms or concepts in text are identified and mapped to entries in domain databases or other resources. Named entity recognition is the identification and classification of named entities (such as persons, locations, or proteins). Relation and event extraction concerns the identification of relationships between named entities and the formulation of this information in a structured manner. Together, named entity recognition and the other extraction tasks are called information extraction. Text summarisation is a process of compressing the essential content of a document into a short passage, whereas sentiment analysis measures the tone of a document, or the attitude expressed by its author, as a whole or with respect to a specific topic. (Ananiadou and McNaught, 2006; Aggarwal and Zhai, 2012; Clark et al., 2013; Cohen and Demner-Fushman, 2014). This thesis concerns information extraction, text summarisation, and sentiment analysis.

In information extraction, the pertinence of entity and relationship types depends on the application at hand and the extracted information may be represented using various formalisms (such as predicates or graphs). There are consequently a variety of representation schemes even within a single domain. Relation and event extraction tasks are typically preceded by named entity recognition, but they can also be solved simultaneously (Bekoulis et al., 2018). Information extraction systems may also include named entity normalisation, which assigns unique identifiers to the extracted entities, facilitating the construction of domain knowledge databases.

Information extraction is similar to semantic role labelling because they share the objective of extracting semantic relationships present in the text. Semantic role labelling focuses on the predicates of sentences and the roles the other syntactic constituents have with those predicates. It is not limited to named entities. In contrast, information extraction is not restricted to any specific syntactic structure[1], but it only concerns the predefined types of entities and relationships. The representation schemes in information extraction often contain formal or informal definitions of those entity and relationship types. Information extraction also emphasises the normalisation of named entities and the construction of domain databases. (Clark et al., 2013).

Biomedical sciences study biological processes and their components, ranging from biomolecules to complex biological systems and diseases, with a focus on human health and provides insights and tools for clinical medicine. The scientific literature in the biomedical field is enormous in size and the number of articles published per year is steadily growing. Researchers have little possibilities to keep track of the newest discoveries even if they fo-

[1] Relevant information may be stated within noun phrases, for example.

cus on very narrow topics. There are many databases that store and provide convenient access to biological knowledge published in the scientific literature but they tend to lag behind the literature due to the need for the manual curation of the knowledge. To solve this information management problem, NLP methods have been developed to assist researchers and curators. Due to its importance, biomedical natural language processing (BioNLP) has emerged as the subfield of NLP that focuses on the biomedical literature. (Cohen and Demner-Fushman, 2014).

In contrast to biomedical sciences, health sciences emphasise the clinical and societal perspectives of human well-being. The scope is not limited to medicine but also include nursing science and public health. A challenge in the healthcare domain is the effective utilisation of unstructured textual data that is produced in clinical environments. Physicians and nurses routinely produce written patient record documents. The volume of patient records tends to be large but the information systems used to access the records suboptimal in design. This hinders professionals' efforts to obtain a clear understanding of the progress of the patient's condition and treatment. Text mining can address these problems by extracting, highlighting, or summarising the relevant information. Textual patient records also contain details that cannot be easily expressed numerically, which makes the analysis of patient record texts a viable option to gain insights into what details are commonly documented, how they are expressed and how they reflect the patient's condition and treatment. Such information may be utilised to develop clinical best practices and to improve the quality of care. Another topic studied in health sciences is patient education, which is often shadowed by medical topics but is nevertheless essential to the success of care. From the perspective of knowledge representation and text mining, written patient education documents are of particular interest.

Political science is a field that aims at understanding how politics influences the development of societies and their relations. Its research topics include political behaviour and decision-making, international relations, and the distribution of power. It is a broad field as politics penetrates all levels of society, from individual to governmental and from domestic to international. (Praag, 2017). Political science studies traditionally employ statistical methods to model and evaluate political phenomena through economic, social, or other variables (Pennings et al., 2006). Information technological approaches have recently gained popularity as a means of quantitatively analysing large-scale textual data that could have previously been explored only qualitatively on a small scale (Kaal et al., 2014). Textual data sets have been recognised as potential cost-efficient sources of information that can complement surveys or other data collection methods of social sciences. Text mining research topics are diverse in nature including, for example, the exploration of political agendas expressed in political speeches and the quantification of attitude in social media postings or news paper articles.

1.1 Research questions and contributions

This thesis contributes to knowledge representation and text mining in three domains: biomedical domain, healthcare domain, and political domain. Its main contributions are six original publications published in peer-reviewed journals and conference proceedings. They address three broad research questions:

1. how can biomedical knowledge be extracted from the scientific literature and utilised in further analyses,

2. how can knowledge representation and text mining be applied to patient records and patient education documents in order to improve the quality of health care, and

3. how can large-scale corpora of newspaper articles be analysed to reveal phenomena that are of interest in political science?

The first research question is addressed by Publications I – III. The first two publications focus on biomolecular event extraction, or the event extraction that concerns named biomolecules (such as proteins and genes) and their interactions (such as binding and regulation). This extraction task has become a major task in BioNLP, which is unsurprising given the importance of understanding biomolecular events as components of biological systems. Publication I studies the transformation of a language-oriented representation of complex biomolecular events into an application-oriented representation of binary relationships, whereas Publication II discusses a component of an event extraction system that leverages syntactic dependency parses. Both publications use the biomedical scientific literature as source data. In contrast, a more generic perspective is taken in Publication III, which focuses on the cross-validation of models learned from pair-input data. Such data is encountered across domains when considering the properties of object pairs, such as the binding of two proteins in the biomedical domain. The study was conducted with protein sequences, rather than textual data, but its results are applicable to text mining because the principles of cross-validation are universal. The study is also a step towards the use of domain knowledge in the design of cross-validation schemes that are more appropriate in specific situations than conventional schemes.

The second research question is covered by Publications IV and V. Publication IV discusses the development of an ontology to model the topics,

Table 1.1: *The thesis contributions (I–VI) cover two research themes and span across three domains.*

Domain	Contribution to knowledge representation	Contribution to text mining
Biomedical sciences	I, III	II
Health sciences	IV	V
Political science		VI

content, and uses of written patient education documents. Patient education documents have been scarcely studied from the knowledge representation and text mining perspectives despite their importance in health care and the possibilities of text mining to improve their quality. The purpose of the ontology is to facilitate the organisation of patient education documents and the development of text mining methods to improve the content of the documents. Publication V explores methods to automatically summarise the text content of patient records to alleviate the information management issues faced by the professionals who need to scan through patients' medical histories.

The last research question is explored in Publication VI, which applies sentiment analysis to assess whether the Western mainstream media has a negative bias towards China, a hypothesis that is being debated by political scientists. The negative portrayal of China has been observed by many previous studies and, building on this observation, Chinese officials have been engaged in changing the Western perception of China from negative to positive. The approach taken in Publication VI uses large-scale corpora of newspaper articles and consequently differs from the previous studies in which small numbers of manually annotated documents were analysed.

Table 1.1 illustrates the distribution of the contributions across the domains and the two research themes. The contributions are mainly in the biomedical and healthcare domains and focus on the application and adaptation of methods to these domains.

1.2 Structure of thesis

This thesis consists of two parts. Part I introduces biomedical sciences, health sciences, and political science as the domains of interest, discusses knowledge representation and text mining tasks in these domains and summarises the contributions in the original publications. Part II contains the reprinted original publications.

In Part I, Chapter 2 discusses the characteristics of natural language and human knowledge in the three domains. It elaborates what kinds of knowledge are typically conveyed by texts in these domains and in which manner such knowledge is expressed. It also highlights the effects of domain-specific aspects and human factors on the nature and expression of knowledge and therefore argues for the need of different approaches in knowledge representation and text mining between the domains. Chapter 3 summarises the six original publications, and the final remarks are given in Chapter 4.

Chapter 2

Domain-specific nature of text and knowledge

The three domains pertaining to this thesis and the associated text resources are vastly distinct in their nature. Two major reasons for their unique properties are the subject matter (and knowledge) of the domains and the objectives of the discourse. The subject matter influences what kind of knowledge is of interest and how it should be represented. For example, biomedical sciences, and to certain degree health sciences as well, study topics where knowledge can commonly be characterised with equations, rules, and other constructs used in natural sciences. In contrast, political science explores concepts that are challenging to define and analyse with the same amount of precision. The text resources also have distinct purposes, which influences their information content and analysis. For example, scientific articles are intended as precise reports of careful investigation whereas patient records consists of possibly brief observations and interpretations of patients' conditions by clinical professionals.

Methodological choices are influenced by both the objectives of the study and the domain-specific needs but the unique characteristics of the domains do not necessarily lead to a need to use unique methods to analyse text and to represent knowledge. Rather, many methods can be used across domains provided that they are adapted to the domain. For example, the overall process of annotating text is the same regardless of the domain but the annotation scheme and its details depend on what kind of information is of interest. The information extraction from textual data involves the same aspects and methodological components across domains but the methods must be adapted to take into account the domain-specific properties of the data. On the other hand, some approaches to discover new knowledge are more suitable for one domain than another. Sentiment analysis, for example, is more useful in analysing political agendas than biomedical topics.

This chapter discusses the characteristics of the biomedical, healthcare, and political domains. The properties of textual data are highlighted and contrasted. The types of resources relevant to this thesis are introduced in the ascending order of unambiguity, starting from the non-text resources used in the biomedical domain.

2.1 Biomedical domain

The biomedical domain combines molecular biosciences (such as biochemistry, molecular biology, and systems biology) with medical sciences and concerns biochemical, biological, and medical topics in the context of human health. Among others, the topics include biomolecules and their interactions, cellular functions and behaviour, as well as diseases, drugs and their interactions. Much of the knowledge in biomedical sciences (such as bioinformatics) builds on the laws of physics and chemistry. Chemical reactions are good examples of phenomena that can be compressed into simple and elegant statements: the behaviour of molecules can be described with reaction equations and their dynamics with rate laws. Moving from chemistry to biochemistry, the size and complexity of the system grows but the underlying chemical mechanisms remain. There are a limited number of biomolecule types in biological systems, of which genes[1], RNA molecules, and proteins are the most important, but the possibilities of small modifications are diverse. The behaviour of proteins may be altered, for example, by phosphorylating, methylating, or acetylating them at specific positions of their amino acid chains. The types of interactions between biomolecules are also limited. The fundamental interaction type is the physical interaction, or binding, as reactions can only occur through interacting atoms, but the physical interactions may be further classified by the effect they produce. For example, a kinase (a protein that catalyses a phosphorylation reaction) is said to phosphorylate its target molecules. In addition to the interactions that are straightforward chemical reactions, indirect functional interactions are of interest for bioscientists because they abstract away from the unnecessary details to the level that is more appropriate for describing cellular processes. For example, a protein may regulate the expression level of a gene through a complex mechanism. The mechanism may be composed of a series direct interactions of several molecules but only the regulation effect is relevant. As a result, both biomolecules and their interactions can be characterised by their inherent properties and modelled using mathematical methods.

[1] In fact, the gene is an abstract concept that is not easy to define (Portin and Wilkins, 2017).

The precise nature of the biomedical domain is reflected in the knowledge representations and databases. They capture biomedical knowledge in a manner that helps researchers to handle the masses of information. Gene Ontology (Gene Ontology Consortium et al., 2000; Gene Ontology Consortium, 2017) is the *de facto* ontology for the characterisation of proteins and RNA molecules. It concerns three aspects of biological knowledge – molecular functions, biological processes, and cellular locations – and associates entities with each other using several types of relations (such as *is a*, *part of*, *regulates*, and *occurs in*). The most detailed entities are very specific (such *negative regulation of pancreatic juice secretion*, *6-hydroxy-3-succinoylpyridine hydrolase activity*, or *endosome to plasma membrane transport vesicle*). Gene Ontology is intended for detailed characterisation of proteins through annotation, and it also supports formal reasoning. Gene Ontology annotations are available, for example, in the UniProt database (UniProt Consortium, 2017), which also contains protein sequences and other information. In general, the content of biomedical resources is manually curated but may contain experimental errors. There may also be algorithmic predictions from non-textual data, which are less reliable than the curated content.

Since the interactions between biomolecules are the underlying mechanism for biological functions, there are many knowledge resources that focus on the characterisation and enumeration of biomolecular interactions. The number of interaction types is typically small. The interactions are often modelled as binary relations, which is a reasonable approach as chemical reactions commonly occur between exactly two molecules. Such a representation scheme is convenient as it results in a graph and facilitates the use of graph theoretical methods to analyse complex systems as interaction networks. Protein–protein interaction databases, such as STRING (Szklarczyk et al., 2017) and IntAct (Orchard et al., 2014), focus on modelling binding, regulation, and post-translational modifications, whereas regulatory interaction databases capture the interaction networks of molecules involved in the regulation of gene expression, such as transcription factors, miRNA molecules, and genes.

Binary relations are not sufficient to fully describe the details of biomolecular interactions. For example, the biological activity of proteins may arise from the formation of an n-ary complex and the different compositions of the complex may exhibit distinct levels of activity. There are efforts, such as IntAct Complex Portal (Meldal et al., 2015) and PCDq (Kikugawa et al., 2012), to address the diverse nature of protein complexes. Protein regulation may also have details that are omitted in binary relations, such as the specific post-translational modification that effects the regulation. There are specialised databases, such as PhosphoSitePlus (Hornbeck et al., 2015), for such details. The networks of binary relations and the detailed descriptions of interactions are both needed: the former omit the details that may only be a nuisance in large-scale biological analyses, whereas the latter support the understanding of the behaviour of individual biomolecules.

In addition to the ontologies and databases focused on the aforementioned phenomena, there are a variety of resources for other biomedical topics. Among others, Cell Line Ontology (Sarntivijai et al., 2014), Drug Target Ontology (Lin et al., 2017), Human Phenotype Ontology (Köhler et al., 2019), and Environment Ontology (Buttigieg et al., 2013) are ontologies that model cell lines, drugs, phenotypes, and environments, respectively. While it could be extended to these topics, the discussion in this thesis is restricted to protein–protein interactions and related molecular-level phenomena. Highlighting the overlap between the biomedical and healthcare domains, SNOMED CT (Wang et al., 2002) is a clinically-focused ontology that also contains molecular-level entities.

Despite the large number of domain databases, the biomedical scientific literature remains an essential source of information regarding new discoveries. The literature contains statements about the same types of entities and processes as the databases. Its crucial difference to the databases is that the knowledge is expressed in natural language. While the scientific literature strives for unambiguous communication, it is not as exact as the databases are because its content is primarily meant for humans. From the perspective of knowledge representation and discovery, the use of natural language introduces a layer of ambiguity which is not detrimental for readers but needs to be addressed in order to transform textual data into structured form.

The extent of ambiguity in the biomedical scientific literature is small in comparison to the other textual sources of this thesis, however, because the scientific literature aims at clear communication and because biomedical knowledge can be expressed with a limited and well-defined vocabulary. The ambiguity mostly stems from the linguistic aspects of written language as a means of human communication. For example, biomolecules may be referred to by their human-friendly names instead of unique database ids, pieces of information may be omitted as they would be obvious for the reader in the given context, or anaphoric expressions may be employed to improve the fluency of the text. As an example, consider the sentence

> *The two nuclear proteins bind to STAT5A target promoters containing GAS thereby stimulating expression of STAT5A regulated milk-genes including β-casein and whey acidic protein.*
>
> (Williams et al., 2004)

which discusses the regulatory effect of ERBB4, or rather of its intracellular domain 4ICD, on the expression of particular genes. The phrase “the two nuclear proteins” refers to the 4ICD domain and STAT5A, the automatic extraction of which is not a trivial task. The mapping of these biomolecules to their database entries requires the knowledge of the species, which must be acquired from the other parts of the article. The specific conditions under which the interactions occur may also be mentioned in the article and should be associated with the interactions in the knowledge representation. The example also highlights the nature of binary interactions: neither “ERBB4 up-regulates β-casein” nor “STAT5A up-regulates β-casein” fully capture the presented knowledge because it is the binding of the ERBB4–STAT5A complex to the promoter that leads to the increased expression level of β-casein.

Early studies on the extraction of protein–protein interactions from text focused on binary relationships (e.g., Thomas et al., 2000; Ono et al., 2001). The research efforts were motivated by the potential of information extraction systems in assisting database curation or even in automatically populating interaction databases (Alex et al., 2008). In this context, binary relationships are relationships that exist between exactly two physical entities (such as proteins) and can be expressed as binary relations. For example, the LLL corpus (Nédellec, 2005) considers genic interactions as directed, untyped binary relationships, whereas the AIMed corpus (Bunescu et al., 2005) uses undirected, untyped binary relationships to capture protein–protein interactions. Similar approaches are used, for example, by TTRUST v2 (Han et al., 2018), which records interactions between transcription factors and their targets as directed, typed binary relationships, and KEGG PATHWAY (Kanehisa et al., 2019), which models regulatory networks as collections of directed, typed binary relationships that cover, among others, the physical binding and regulation of proteins.

Binary relationships are problematic as a representation of the knowledge conveyed by the scientific literature because the text often contains information that cannot be reduced to typed binary relationships. In particular, the statements of one physical entity influencing a relationship between other physical entities cannot be expressed as binary relationships. Relationship details, such as experimental conditions, subcellular locations, or hedging, are also challenging to include into the representations that are based on binary relationships. Such information may be implicitly covered by the scope of the resource, included as additional attributes, or simply ignored. For example, the ComPPI database (Veres et al., 2015) uses species, subcellular location, and experimental evidence as attributes but ignores temporal aspects (such as the phase of the cell cycle and the developmental stage of the organism). From the perspective of BioNLP research, the extraction of typed binary relationships is an interesting task but additional tasks, such as the extraction of supplementary information and the mapping of gene and protein mentions to database identifiers, also need to be considered to fully support biomedical research via database construction.

It was observed during the development of BioInfer[2] (Pyysalo et al., 2007) that complex relationships can be used to conveniently capture the details of the relationships expressed in text. Here, a complex relationship refers to a construct that involves more than two physical entities or more than one relation. The key difference between binary and complex relationships is the manner in which the interactions of entities are expressed. To take a graph theoretical view, a binary relationship can be represented as two nodes and an edge between them. The nodes represent physical entities and the information regarding their interaction is encoded to the edge.[3] A complex relationship cannot be expressed with only one edge. Instead, the information on the interaction(s) must be encoded to the combinations of nodes and edges. This results in a directed acyclic subgraph in which both physical entities and the instances of interactions are represented by nodes. The edges represent the roles of the entities in the interactions. Naturally, binary relationships can also be expressed in this form but the extra node and the edges do not provide any additional information compared to the single directed, typed edge.

Binary relationships are entity-centric and emphasise the network of interactions, whereas complex relationships can express interactions independently from entities, which is closer to the syntactic representations used in BioNLP. For example, the syntactic structures under the basic version of the Stanford typed dependencies representation (Marneffe and Manning, 2008b) are directed acyclic graphs (Marneffe and Manning, 2008a). The complex relationships of BioInfer indeed mostly follow Stanford dependency structures (Björne et al., 2008).

Complex relationships emphasise the details of interactions and resemble the approach taken in semantic parsing, which concerns the discovery of the participants of events, actions, or other activities along with their roles (Reddy et al., 2017). The similarities and differences between binary and complex relationships as well as dependency parses are illustrated in Figure 2.1.

Complex relationships can have other relationships as participants, which is not possible in binary relationships. This makes complex relationships more versatile than binary relationships in representing biomedical knowledge but also more challenging to analyse. The regulation of biomolecular and cellular activities is a major phenomenon and regulatory relationships may have relationships as participants. For this reason, some bioinformatics databases, such as the Reactome knowledgebase (Fabregat et al., 2018), also use representations in which such relationships can be expressed.

[2] BioInfer is currently available at `http://mars.cs.utu.fi/BioInfer/`.

[3] In theory, the edge can encode any amount of information but, in practice, the set of relationship types is small to ensure the occurrences of each type are sufficiently abundant.

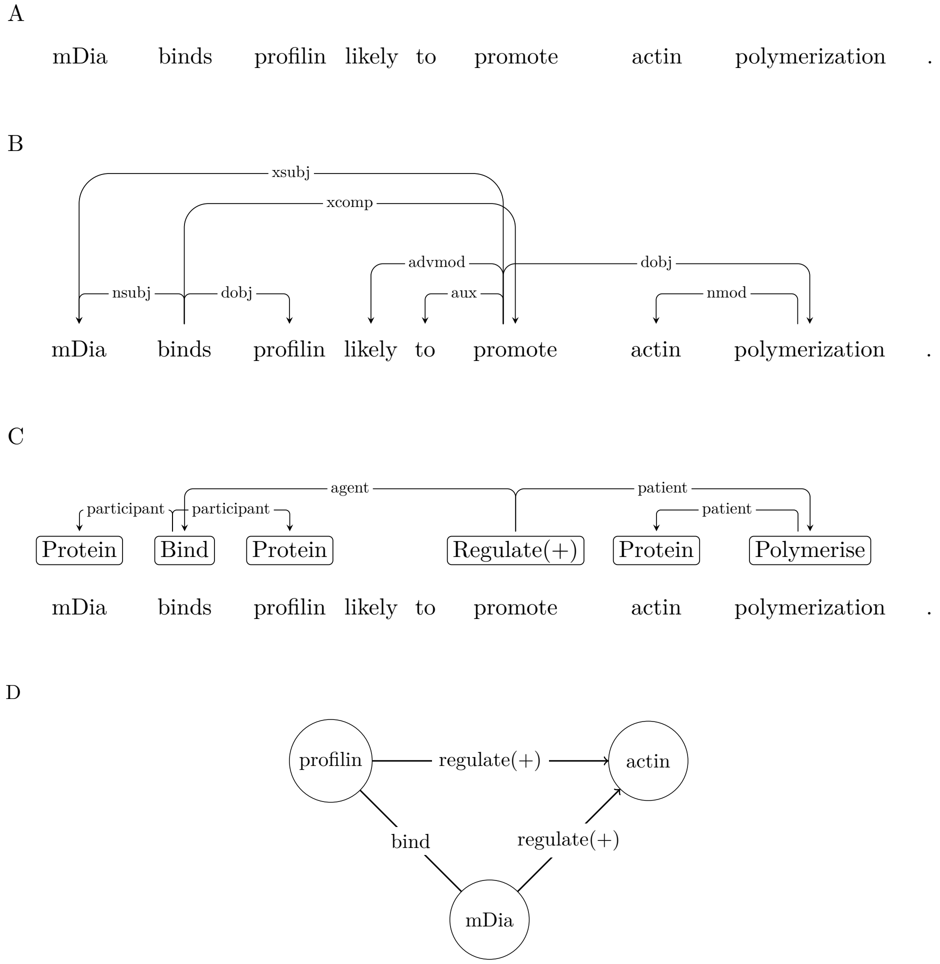

A
mDia binds profilin likely to promote actin polymerization .
B
xsubj
xcomp
advmod
dobj
nsubj
dobj
aux
nmod
mDia binds profilin likely to promote actin polymerization .
C
agent
patient
participant
participant
patient
Protein
Bind
Protein
Regulate(+)
Protein
Polymerise
mDia binds profilin likely to promote actin polymerization .
D
profilin
regulate(+)
actin
bind
regulate(+)
mDia

Figure 2.1: *A) A sentence contains detailed information on protein–protein interactions. B) The Stanford dependency parse of the sentence is a directed acyclic graph. C) Complex relationships capture the details of the stated interactions. The existence of an interaction is represented by a node, which leads to a directed acyclic graph. D) Binary relationships capture the essential pieces of information on the interactions. An entity-centric network of interactions is formed.*

The discussion above illustrates the spectrum of knowledge representations used in capturing biomolecular interactions. On one hand, there is a strong biological motivation to use binary relationships because they are useful in network analysis. On the other hand, complex relationships resemble dependency parses and are able to include the details of interactions, which justifies their use from the linguistic perspective. BioNLP and bioinformatics use both types of knowledge representations, yet their focus is on the linguistic and biological aspects of the biomedical domain, respectively.

The objectives of the research on the acquisition of biomedical knowledge from text include not only developing methods to transform text into structured form but also discovering solutions to overcome the imprecisions introduced to otherwise precise biological knowledge by the nature of human communication. Since the binary interaction networks used by bioscientists differ from the complex relationships convenient in expressing statements in text, there is a need to bridge the gap between these two representations.

2.2 Healthcare domain

Electronic patient records have the same purpose as their paper-and-pen predecessors: to document the professionals' observations and conclusions regarding a patient's condition and record the progress of the subsequent treatment. Despite the possibilities of structured data, much of the information in electronic patient records is in free text because it is challenging to design schemes which are able to address the diversity of the content of those records. (Häyrinen et al., 2008; Evans, 2016). Among others, electronic patient records contain physicians' notes on patient encounters, admission and discharge notes regarding patients' stays in a hospital, nurses' notes on care given in a ward, prescribed and administered medications, patients' diagnoses, laboratory test results, and radiologists' reports. Diagnoses, procedures, medication and laboratory test results are examples of information that is commonly stored in structured form. (Häyrinen et al., 2008).

From the perspectives of content and language, electronic patient records are markedly different from the biomedical scientific literature in several aspects. First, the focus is on individuals and their health rather than on the molecular-level mechanisms of diseases (Friedman et al., 2002). Second, patient records are documentative and hence include observations that in research would be considered data for further analysis rather than results to be reported in a scientific article (Evans, 2016). Third, the language used in patient records is not as linguistically refined and formal as in the scientific literature (Friedman et al., 2002; Laippala et al., 2009). As an example, consider the passage

> *A 84-year-old man comng to emergency. Mild chest pain, lasted for two days. Pale skin butt no signs of physical injury. No history of angina pectoris.*

which is a hypothetical documentation of the arrival of a patient. It contains background information ("84-year-old man" and "no history of angina pectoris"), patient's own description of his condition ("mild chest pain") and its details ("lasted for two days"), and observations by the physician ("pale skin" and "no signs of physical injury"). Note that the patient's and physician's observations are mixed into the same passage. The passage contains phrases rather than full sentences, and there are two spelling mistakes ("comng" and "butt").

Ideally, patient records contain comprehensive and fully accurate descriptions of the actual state of the patient at the time of recording. In practise, however, the records are only the best current knowledge based the domain expert's observations and conclusions. (Lenert, 2016). Diagnostic uncertainty is inherently present in patient records because professionals have limited and imperfect knowledge of the patient (Bhise et al., 2018). Some diagnoses, such as fractured bones, can be recorded with high certainty but others, such as mental health problems, have more uncertainty involved in them. Suspicions of diseases may not be confirmed until new pieces of information become available through further examination. Patient records are hence not as refined and complete as scientific articles and tend to have higher level of uncertainty than scientific texts reporting the results of lengthy analyses.

In addition to the varying uncertainty, the degree of interpretations made by professionals varies because some professionals are more confident than others to record their conclusions regarding their patients' health conditions. Zalon et al. (2017) observed that delirium is often not mentioned in patient records although indicative symptoms are, which is consistent with the observation by Numan et al. (2017) that making a delirium diagnosis is difficult. There is also variability in the documented topics and in the vocabulary used to document them (Considine et al., 2016; Boyd et al., 2018; Gonçalves et al., 2019). As an extreme example, the sentence

The patient acted slightly confused in the morning.

explicitly expresses the confusion of the patient as interpreted by the nurse, whereas the sentence

The patient sees flying elephants outside the window.

repeats the experience the patient communicated to the nurse as a fact and only implicitly states the confusion.

The third factor influencing the nature of knowledge in patient records is the fact that some pieces of information originate directly from the professionals while others are expressed by the patient and interpreted by the professionals (Lenert, 2016). For example, professionals can estimate the amount of pain the patient experiences by observing indicators such as facial expressions, body movements, and compliance with ventilation (Pudas-Tähkä et al., 2014) but the patient can also communicate the pain verbally to the professional or use a pain measurement tool (such as Numeric Rating

Scale) (Karcioglu et al., 2018). However, the sensitivity to and the expression of pain vary between patients (Giordano et al., 2010), and patients may even adjust the level of pain they report in anticipation of a certain response by professionals (Dijk et al., 2016). Given the subjective nature of many themes relevant in health care, it is challenging, if not impossible, for patient records to convey the exact knowledge and be consistent across documents.

Regarding all the aforementioned factors, healthcare professionals are trained to pay attention to the relevant details in order to produce consistent records throughout the care but the human factor is nevertheless influencing the degree of uncertainty and interpretation contained in the records. The variation in recording details leads to vocabularies that tend to be larger than the vocabularies needed to express knowledge in the biomolecular domain. This poses a challenge for information extraction methods.

The incomplete knowledge, the documentative approach, and the need for interpretations, together with the complexity of the biological and societal processes pertaining to health conditions and care, result in textual data that has more uncertainty and ambiguity than the scientific literature. The responsibility is partially given to the reader (or the computational analysis) to make conclusions regarding the significance of the observations. Because of this inherent ambiguity, some degree of uncertainty must be accepted in knowledge representations in healthcare settings. The aforementioned characteristics of patient records are also reflected in the experimental setups that rely on manual content analysis, possibly aided by text mining or knowledge representation, to discover new domain knowledge (see, e.g., Sjöblom et al., 2013; Kauhanen et al., 2014; Uronen et al., 2017).

Patient education documents are another type of documents commonly encountered in healthcare settings. They concern similar topics as patient records but they are intended to provide information for patients instead of recording their care. They discuss topics that patients need to know about their health conditions or care, give instructions on actions that patients need to take, and are authored for laymen in terms of appearance and textual style. (Johansson et al., 2004). As such, their style and lack of inherent uncertainties bring them closer to regular text than patient records but their content cannot be regarded as precise as biomedical scientific articles.

The scope of the documentation produced in healthcare settings is broad. The topics that need to be addressed by text documents and knowledge resources include, among others, human anatomy and physiology; health conditions, such as diseases; causes of health conditions, such as bacteria and accidents; consequences of health conditions, such as symptoms; physical environments, such as hospitals; organisational structures, such as medical specialities; and care-related activities, such as medical procedures and rehabilitation.

Knowledge resources in the healthcare domain tend to focus on the medical and nursing perspectives of health care because they are needed to create consistent and interoperable records. For example, the Finnish versions of the International Classification of Diseases (Komulainen, 2011) and the International Classification of Primary Care (Kvist and Savolainen, 2010), the THL Classification of Procedures (Lehtonen et al., 2013), and the Finnish Care Classification (Liljamo et al., 2012) are widely used in Finnish healthcare environments to document diagnoses, patient encounters, clinical procedures, and nursing care. As such, they are primarily intended for documentation rather than formal representation of healthcare knowledge.

There are also theoretically and formally motivated resources, such as the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2008) and SNOMED CT. The FMA is a comprehensive formal ontology of human anatomy designed to be a reference ontology that can be reused in application ontologies. It contains over 75.000 types of entities and over 190 kinds of relations. (Rosse and Mejino, 2008) In contrast, SNOMED CT is a comprehensive resource for communicating patients' characteristics, their backgrounds, and their care in patient records and other documents in a systematic and interoperable manner. It covers a wide range of entities. Starting from molecular-level concepts, such as drugs and metabolic products, and cell-level concepts, such as micro-organisms and cell types, it overlaps with resources used in the biomedical domain but its focus is on the clinical use of the concepts. Recent efforts have focused on improving the semantics and interoperability of SNOMED CT (Rodrigues et al., 2015; Schulz and Martinez-Costa, 2015; El-Sappagh et al., 2018).

Patient education and patient education documents have received little attention with respect to knowledge representation. Patient education related concepts are present in SNOMED CT and Logical Observation Identifier Names and Codes (LOINC) (McDonald et al., 2003) but the details of the content of patient education documents are not modelled by them. In the field of nursing science, patient education documents have been studied to discover what properties are needed to make them understandable and informative, and some content topics have been identified. The discoveries were not added to knowledge resources in these studies, however.

2.3 Political domain

Political science studies the behaviour of individuals, institutions, and societies from the perspective of politics, which includes topics such as political power, public policy, political decision making, and international relations (McAuley, 2003; Wilkinson, 2007). The analysis of textual documents is a commonly used method in political science, and recent advances in electronic data sources and computational methods have promoted the use of automated quantitative analysis (Wilkerson and Casas, 2017). Computational methods enable large-scale analyses that cannot be achieved by the manual examination of text documents. Among others, text mining has been applied to news articles (Fortuny et al., 2012), senatorial speeches (Diermeier et al., 2011), and social media posts (Driscoll and Thorson, 2015). At its simplest, text documents can be classified as a whole by the themes they discuss, such as economy or sports. However, such classification does not provide the detailed knowledge of content that would be beneficial in political behaviour studies.

News and social media texts are markedly different from scientific articles and patient records. They cover a multitude of topics, commonly concerning societal or political matters, and linguistic styles. There are hence many opportunities to perform specialised analyses on linguistic phenomena such as rhetoric (Zhang et al., 2017), analogies (Lofi et al., 2014), and satire (Rubin et al., 2016) that are relevant in the political domain. Automated text analysis can be performed by applying well-established methods, such as word embeddings and artificial neural networks (Rudkowsky et al., 2018) or dictionary methods (Young and Soroka, 2012), to appropriately formulated research questions.

Local and global events shape our world and influence the attitudes of individuals. The analysis of events is thus useful in political science. The extraction of named entities (such as persons and organisations) and events from text facilitates large-scale analyses and provides opportunities to discover non-trivial connections between entities as well as between events. Given the diversity of the topics covered by news and social media, it is a challenge to model pertinent entities and events at the same level of details as in the biomedical and healthcare domains unless only a specific subset of events is considered.

The Conflict and Mediation Event Observations (CAMEO) is a classification of conflict-related activities (Gerner et al., 2002). It focuses on events between groups of people or institutional bodies, such as *Demand ceasefire* and *Threaten blockade*. It is used in the Global Data on Events, Location and Tone (GDELT) datasets, produced by an ongoing project to record the entities and events appearing in the news media around the world (Leetaru and Schrodt, 2013). The newest version of GDELT (v2.1) uses the Global Content Analysis Measures system (GDELT Project, 2015), which extends the scope beyond the CAMEO events by extracting emotional and thematic information with several dictionaries. Together with the inclusion of multiple news sources and languages, it makes the GDELT datasets very comprehensive knowledge resources focusing on news media. There are also recent research efforts to construct formal ontologies that can capture the logical connections between events and hence facilitate more detailed analyses (Brown et al., 2017; Segers et al., 2017).

Rather than being purely documentary, news articles and social media posts are written from a specific perspective to communicate a story, an idea, or an opinion. They are also motivated by the desire of the author to participate in public discussion and as such they may be intended to shape the views of others. Entman (2007) argues that two types of bias should be considered. The content bias occurs when the framing of a matter consistently favours one of the two opposing sides over the other over time. Framing refers to the process of constructing a narrative that highlights a perspective to the matter. It can be used to direct readers to accept or support a particular stance. As an example of the content bias, a report of a protest may ignore the events that preceded the protest and only emphasise its adverse consequences. The decision-making bias concerns the factors that influence the choices the authors make regarding what to report and how to frame the matter. Such factors include, among others, personal ideology, lobbying by stakeholders, and the occurrence of other news-worthy events at the same time. For example, Field et al. (2018) observed that a decline in the Russian economy leads to the state-controlled media in Russia to shift their coverage towards news about the United States of America.

The quantification of a media bias is a challenging task because the analysis must properly model the influence of text on the thoughts and actions of readers (Entman, 2007). The influence is effected not only by the framing but also by the stance and sentiment of the text. Sentiment analysis is hence an essential task in political science (Ahmad et al., 2011; Abercrombie and Batista-Navarro, 2018; Bhatia and P, 2018). A common approach is to detect the polarity of the text, which only considers whether the text conveys a positive, negative, or neutral message, but the analysis may also focus on the intensity of the sentiment (Tian et al., 2018). Sentiment dictionaries, such as Lexicoder Sentiment Dictionary (Young and Soroka, 2012), and dictionary-based methods have a significant role in the domain.

In addition to the aforementioned types of biases, news or social media texts can simply represent the reality in a distorted manner. Peaceful protesters may be unwarrantedly portrayed as a violent mob, for example. Entman (2007) argues that “almost any nontrivial reality will be controversial” and that the analyses should hence focus on the content and decision-making biases. However, the possibility of distortion should be taken into account by the information extraction methods. On one hand, if they are intended to reflect the interpretation to which the readers are exposed, named entities and events should be extracted as they are stated in the text. Subsequent analyses can then explore how the discourse is used to further political agendas. On the other hand, if an objective account of events is desired, the reliability of the statements must be taken into account to correct the distortions.

In summary, automated text analysis in the political domain must take into account features that are not encountered in the biomedical scientific literature or the text documents written by healthcare professionals. News and social media texts are diverse in content and style, they are intended to shape readers' opinions rather than to be documentary, and the phenomena under study are complex. For example, Zerva and Ananiadou (2018) observed that the detection of uncertainty is more difficult in the newswire domain than in the scientific domain. Some objectives and methods (such as the information extraction of events) are shared between the domains, but some (such as sentiment analysis) are not needed in the analysis of scientific articles or patient records.

Chapter 3

Thesis contributions

3.1 Complex-to-binary mapping of protein–protein relationships

Publication I addresses the gap between the linguistically and biologically motivated approaches to the representation of biomolecular events. Its main purpose is to demonstrate that it is possible to transform the complex relationships that closely describe statements in texts to binary relationships, which are commonly used by the domain experts. The underlying hypothesis is that the complex relationships extracted from a sentence contain more knowledge than the binary relationships extracted from the same sentence. It is hence possible to transform a complex relationship into one or many binary relationships that cover the knowledge that could be directly extracted as binary relationships from the sentence. This transformation process may involve simplification and a loss of knowledge that cannot be expressed as binary relationships.

The study was conducted on the BioInfer corpus because it was the only available corpus containing complex relationships at the time. The knowledge representation of BioInfer also has a wide range of relationship types, which further motivates the use of BioInfer. Some types concern specific biomolecular events, whereas others are about experimental evidence, static relations, or unspecified interactions. The types representing uncertain or incomplete knowledge of interactions facilitate the expression of knowledge as it is stated in the text. Figure 3.1 illustrates the types of relationships annotated in BioInfer. In comparison, the knowledge representation scheme of the GENIA Event corpus (Kim et al., 2008) is similar but strictly focuses on cellular and biomolecular events.

BioInfer uses two taxonomies that describe physical entities and their relationships. The Relationship Ontology defines the relationships that are annotated in the corpus. They include both biomolecular events and static

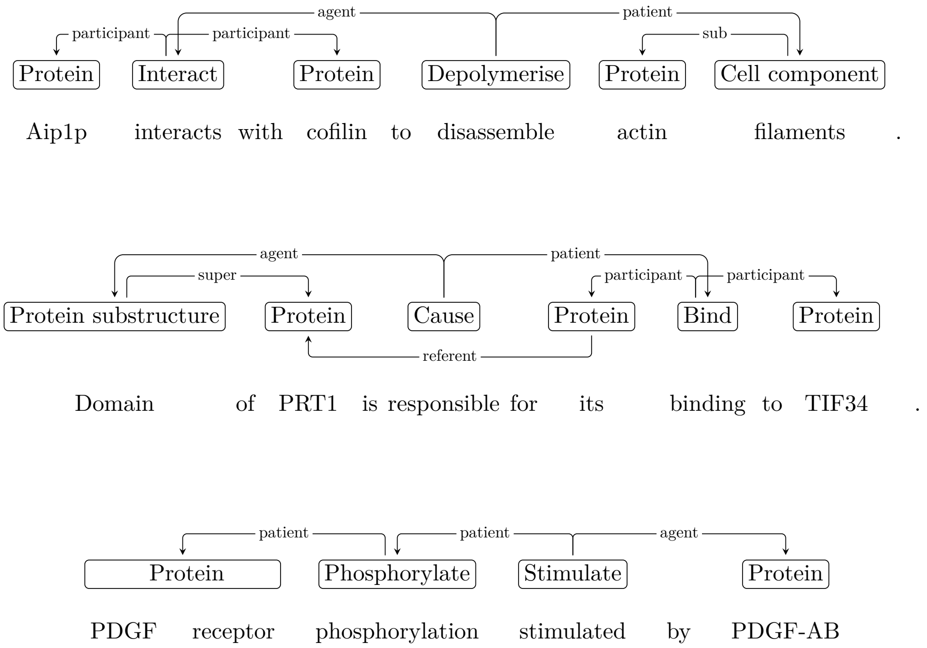


Figure 3.1: *Examples of complex relationships in BioInfer. The relationships are shown as semantic networks. The text bindings of the entities and relationships are indicated by their alignment with the text. For clarity, the names of the entity and relationship types have been simplified and the prepositions have been removed from the text bindings.*

relationships, such as memberships in protein complexes and coreferences. In contrast, the Entity Ontology contains physical entities and their properties but also biomolecular processes because BioInfer entities are units needed to describe relationships, which always involve at least two entities. The taxonomies facilitate the construction of arbitrarily large complex relationships via the nesting of predicates and entities. (Pyysalo et al., 2007).

Nested entities in BioInfer have specific but implicit relationships with each other. The BioInfer annotation was hence normalised into a graph representation in which all associations are explicitly stated as edges. This transformation also unified the representation of biomolecular processes.

Since biomolecular mechanisms are well-characterised and limited in number, a rule-based approach to binarisation was taken. It allows the use of human-knowledge regarding the relationship scheme. The binarisation process is a sequence of transformations that simplify a complex relationship until it becomes binary. This is illustrated in Figure 3.2. The transformation rules were manually designed to first normalise the original structures, in which the same knowledge may have various forms, and then gradually remove the least important pieces of information. Since a single complex relationship may lead to several binary relationships, the rules are exhaustively applied to recover all valid binary relationships. In essence, there are two types of rules: those that remove excess information and those that infer simplified relationships using the properties of the BioInfer relationships.

The Prolog language (Wielemaker et al., 2012) was used because it has the flexibility of the first-order logic and the order in which the rules are applied can be controlled. It is hence a suitable tool to implement a step-wise simplification and approximation process. Since Prolog is a logic programming language, the knowledge and the rules can be conveniently expressed using predicates. A graph-based approach, such as the Resource Description Framework (Cyganiak et al., 2014), could also have been used.

During the development of the binarisation process it was concluded that there are three independent and complementary facets of knowledge that should be preserved. The process facet describes the biomolecular process of the target molecule, such as phosphorylation, that the interaction concerns. For direct interactions, it is the knowledge that is typically recorded in biological databases. The interaction may also be regulatory in its nature, indicated by another facet, in which case the regulatory molecule is indirectly responsible for an effect on the indicated process. Last, the negation facet indicates whether the occurrence or non-occurrence of the mentioned interaction is asserted in the text. In a biological study, the former is of particular interest but text mining research benefits from both.

A simple random sample of 50 sentences that had not been used in the development process was selected for the evaluation of the binarisation quality. The correctness of each relationship was manually checked and, if incorrect, the type of error was recorded. The results suggest that the binarisation process is viable, which can be attributed to the precise properties of the BioInfer relationships. The binarised version of BioInfer was later compared by Pyysalo et al. (2008) to four other corpora containing binary relationships. The proposed approach should generalise to other corpora, given the well-defined nature of biomolecular interactions.

The study provides evidence for the hypothesis that it is a viable approach to first extract as complete complex relationships as possible and then transform them into binary relationships if necessary. The future research should examine the generalisability of the transformation rules across schemes and eventually focus on the development of statistical binarisation methods. A generalised binarisation method would not only facilitate the transformation of biomolecular knowledge to fulfil the needs of applications but also give insights into the ontological nature of protein–protein interactions and help to examine the various approaches to protein–protein interaction extraction, a matter encountered by Kim et al. (2016) during the Genia event extraction task of the BioNLP Shared Task 2016.

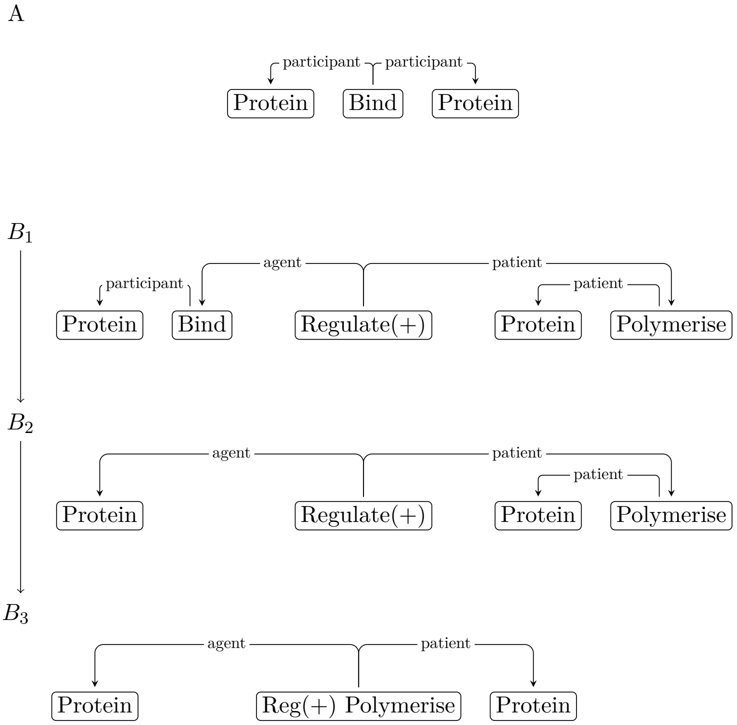


Figure 3.2: *The binarisation process of the relationships stated in the sentence in Figure 2.1. A) The binding between* mDia *and* profilin *is trivially binarised.* B_1*–*B_3*) The relationship between* mDia *and* actin *is simplified in two approximative steps. The* profilin–actin *pair is processed similarly. The relationship type* Reg(+) Polymerise *means that there is a positive regulatory relationship that pertains to the polymerisation of the target protein.*

3.2 Reconstruction of protein–protein relationships

The information extraction of protein–protein interactions has received significant attention in the BioNLP field because of its importance in the curation and expansion of biomedical databases. The popularity of the task was particularly increased by the shared tasks organised by the BioNLP research community. The BioNLP Shared Task 2009 (Kim et al., 2009) focused on the extraction of biomolecular events, and the later shared tasks in 2011 (Kim et al., 2011), 2013 (Nédellec et al., 2013), and 2016 (Nédellec et al., 2016) expanded the scope to epigenetics, cancer genetics, bacteria biotopes, and others.

Björne et al. (2009) approached the extraction of protein–protein interactions as a problem of generating a directed acyclic graph of interactions

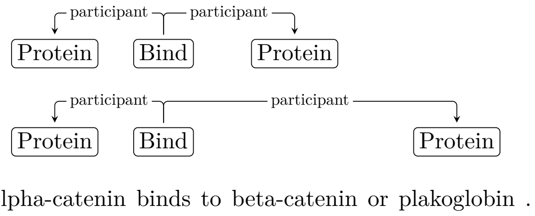


Figure 3.3: *An example of a sentence in which two instances of relationships are expressed by a single word.*

from a dependency parse graph. In this approach, words or phrases are associated with the nodes of the graph, which is essentially a named entity recognition task targeting both physical entities and biomolecular events. The construction of relationships is then achieved by determining the edges of the graph. The approach was found to be very robust in the extraction of protein–protein interactions in the BioNLP Shared Task 2009 and succeeded in the later shared tasks as well (Björne and Salakoski, 2011; Björne and Salakoski, 2013).

In named entity recognition, each entity is associated with its own unique word or phrase in text, which may be nested or overlapping. This approach is not sufficient to extract the relationship nodes of a semantic graph because a single word or phrase may represent multiple instances of relationships. For example, in the sentence

Alpha-catenin binds to beta-catenin or plakoglobin.

the word “binds” represents two separate interactions as can be seen in Figure 3.3. Therefore, some nodes generated by the entity recognition process must be duplicated and their edges distributed properly among the duplicates. The relationship graphs in which the text bindings of the nodes are unique are called projected relationships because they can be considered to be projections into a lower dimensional representation in which some nodes cannot be distinguished from each other.

Since the number of relationship nodes represented by a word depends on the relationships themselves, i.e. the manner in which the edges are connected, separate node and edge extraction steps cannot produce the final relationship graphs. A deprojection step is needed to determine which nodes must be duplicated and how their edges should be organised into the duplicates.

Publication II explores the deprojection task to develop a method to execute the deprojection and to obtain insights into its characteristics. Instead of using brute force to determine which combinations of edges form correct relationships, it leverages dependency parse graphs and the restrictions imposed by relationship types. The deprojection method relates to the idea of collective and distributive readings, which is discussed by Brisson (2003).

The deprojection method has two steps. First, the successors of a relationship node are grouped by their semantic role in the relationship and by the syntactic similarity of the associated words. This process reveals groups in which the members tend to participate either in the same instance of relationship or each independently in their own instances. For example, considering the sentence

A binds to B and C.

the *Bind* event is connected to three participants. At first, they form a single group, but B and C are then separated from A because the positions of their text bindings in the syntax differ from that of A relative to "binds". Last, the groups are classified as distributive or collective nodes, which determines whether the predecessor should be duplicated for each successor and the successors distributed or whether the successors should reside under the same predecessor, respectively. In order to produce the final graph, the duplicates produced by a distributive node are classified and processed similarly. This is illustrated in Figure 3.4.

The experimental results on BioInfer and GENIA Event corpus show an improved performance over the earlier solution by Björne et al. (2011). This suggests that the use of biological and linguistic properties of the relationships benefits the process. In the future, the method should be extended to other semantic roles, such as *located in*, which result in more complicated combinations of successors. The exploration of the representation of mutual exclusion, joint action, simultaneity, and other relevant concepts modelled by formal ontologies may also benefit the further methodological development and help to bring the text mining and ontological knowledge representation approaches closer together. Given the precise semantics of biomolecular processes and the clarity of biomedical scientific texts, in an ideal case, it could be possible to approach the extraction of biomolecular events as a task of constructing a one-to-one mapping of nodes and edges from a syntactic graph to a semantic graph followed by the deprojection process that, leveraging the syntactic and semantic properties, produces a formal representation of the complex relationships.

Since the execution of this study, the deprojection task has remained relevant, given the appearance of complex relationships in several BioNLP shared tasks (Kim et al., 2011; Nédellec et al., 2013; Kim et al., 2016) and the development of deprojection components by Björne and Salakoski (2013) and Björne and Salakoski (2018), who applied support vectors machines and neural networks, respectively, to solve the deprojection task in their information extraction systems.

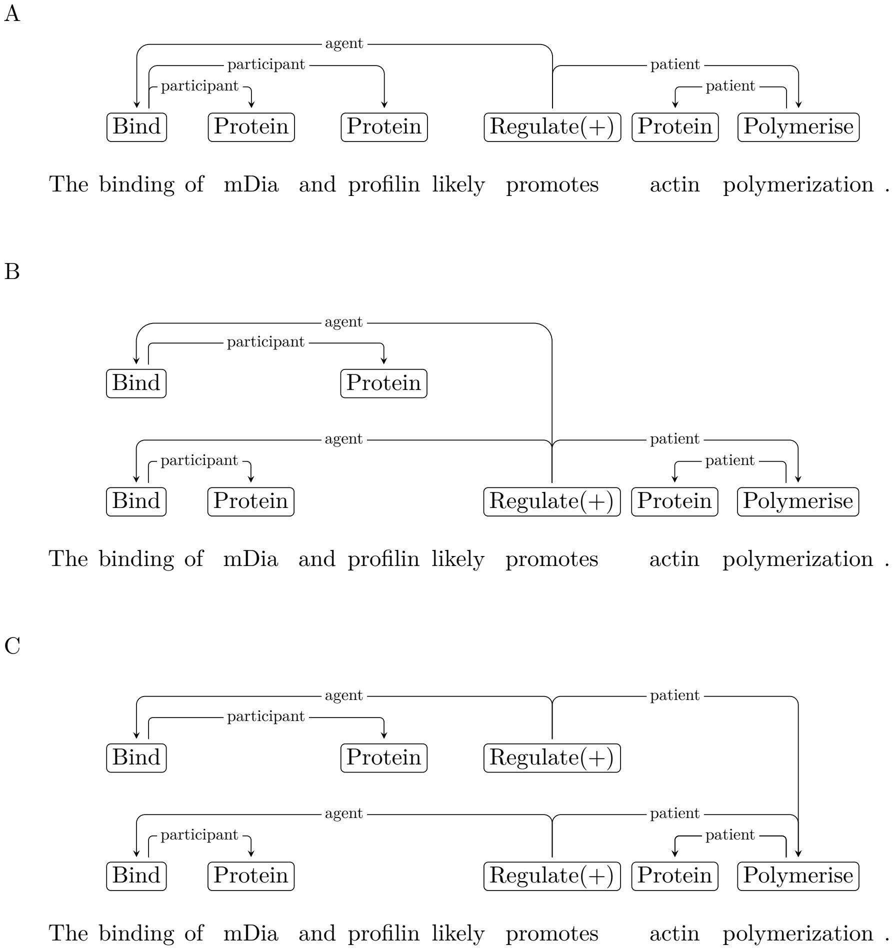


Figure 3.4: *The sequence A–B–C illustrates the deprojection of the complex relationships stated in a sentence rephrased from Figure 2.1. If distributive, the* Bind *node is duplicated for each of its successors (A–B), and subsequently, if also distributive, the* Regulate(+) *node is duplicated for each* agent *edge while associating the* patient *edge to each copy (B–C). Each graph has its own meaning: A)* mDia *and* profilin *bind each other, which triggers the positive effect on the polymerisation. B)* mDia *and* profilin *bind separately to unstated molecules, and these binding events jointly effect the regulation. C) As previous but the binding events regulate the polymerisation independently.*

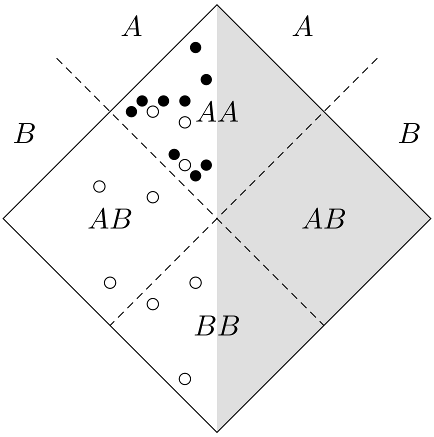
3.3 Cross-validation on pair-input data

The estimation of generalisation performance is an essential step in the evaluation of a learned model. Its objective is to quantify the quality of the model in a situation where the model is applied to previously unseen instances. The estimation is typically performed by randomly partitioning the known instances into training and validation sets. A model is then learned from the training set and evaluated against the validation set. (Cios et al., 2007). The procedure arises from the assumption that the instances are independent from each other and that the consequent independence of the training and validation sets thus reflects the assumed independence of the known and unknown instances. This approach is commonly used in cross-validation schemes, in which the generalisation performance is estimated as the mean of multiple partitioning–training–evaluation rounds. (Arlot and Celisse, 2010).

There are many situations in which the independence assumption does not hold in the biomedical domain. Pair-input data is one of such cases. Pair-input data refers to instances that represent pairs of objects instead of individual objects. For example, in the prediction of protein–protein interactions from their sequences the instances concern protein pairs and contain information from both sequences. An instance that shares a pair member also shares some of that information, which makes the instances dependent on each other. (Park and Marcotte, 2012).

Symmetric pair-input data refers to a situation where the pairs are composed of objects of a single type (e.g. two proteins) and where labels arise from a symmetric relation. Again, dependencies naturally arise from objects being shared between instances. As demonstrated by Park and Marcotte (2012), this has drastic consequences for the estimation of generalisation performance because there are three types of previously unseen pairs in this situation: AA, AB, and BB pairs which contain two, one, or zero objects present in the set of known instances, respectively. This is illustrated in Figure 3.5. Note that an AA-type instance represents a pair of objects that are present in the training data in some other pairs.

In general, the conventional cross-validation approach is not appropriate for pair-input data because of the dependencies (Pahikkala et al., 2012). Park and Marcotte (2012) demonstrated that the conventional cross-validation approach may or may not be appropriate in the protein–protein binding classification task. The validity of the generalisation performance estimate depends on which kinds of pairs the learned model will be applied to. The estimate is acceptable for AA-type instances. However, the estimate is over-optimistic for AB-type instances and even more over-optimistic for BB-type instances. This highlights the need to evaluate models separately for each intended purpose. The latter two cases are encountered in the biomedical domain because new proteins are regularly discovered and their interactions with the previously known proteins as well as with each other are of interest. The tendency of over-optimistic estimates is similar to how the performance of a model tends to be over-estimated if the distribution of the training instances is not the same distribution from which the future



(a) *Object sets A and B lead to three types of instances (AA, AB, BB).*

Known	Unknown		
	AA	AB	BB
$(1,3)$	$(2,8)$	$(3,14)$	$(11,17)$
$(1,8)$	$(4,7)$	$(6,12)$	$(12,14)$
$(1,9)$	$(6,9)$	$(8,18)$	$(16,19)$
$(2,7)$			
$(3,4)$			
$(3,6)$			
$(5,9)$			
$(7,8)$			
$(7,9)$			

(b) *The pairs of objects that correspond to the instances marked in (a).*

Figure 3.5: *Symmetric pair-input instances can be partitioned into three categories according to their compositions with respect to the objects in the known data. In (a), the slanted axes indicate the indices of the objects, partitioned by their presence (A) and absence (B) in the known data. A point in the diamond-shaped area corresponds to an instance derived from a pair of objects, and the instances that are redundant due to symmetry are shaded. The known instances and some examples of the unknown instances are indicated by the black and white points, respectively. In (b), the object compositions of the marked instances are explicitly stated.*

instances will be drawn, although in some cases mismatching distributions may be desirable (González and Abu-Mostafa, 2015). For example, a model learned to predict the structural families of aqueous proteins will very likely fail if applied to membranous proteins, which have very different structures.

Publication III examines the properties of two cross-validation schemes for symmetric pair-input data in the setting where the generalisation performance on *BB*-type instances is of interest. The two schemes, relaxed and strict, involve the partitioning of the known instances at the object level. That is, the objects are first assigned into training and validation groups and the instances are then partitioned according to which groups their objects belong to. The validation set is the same in both schemes and contains those instances than can be formed by pairing the objects in the validation group with each other (i.e. the *BB*-type pairs). The training set in the relaxed scheme contains the remaining instances (i.e. the *AA*-type and *AB*-type pairs). The scheme resembles the conventional cross-validation scheme and, in particular, there are dependencies between the training and validation sets. In contrast, the training set in the strict scheme only contains the instances that can be formed by pairing the objects in the training group with each other (i.e. the AA-type pairs). As a result, there are no dependencies but some instances are in neither of the two sets.

The mean and variance of the deviations of the cross-validation estimates from the estimate of the real performance were experimentally observed using 1000 independent repeats from a large data set. In each repeat, 100 proteins were randomly selected from the set of more than 380,000 proteins. All of their possible pairs were included into the known data with which the training and cross-validation was performed. The real performance of the learned model was estimated with a random sample of 10,000 BB-type instances. The hypothesis was that the relaxed scheme exhibits a large positive mean deviation, following the observations by Park and Marcotte (2012), because it effectively estimates the performance on AA-type instances instead of BB-type instances. The strict scheme was hypothesised to not have such a positive mean deviation because it reflects the setting in which the model will be used (i.e. no dependencies). Instead, a small negative mean deviation was expected because the number of training instances in the cross-validation is smaller than when learning the final model. The variance of the strict scheme was also expected to be larger than that of the relaxed scheme because the training sets are smaller in the strict scheme.

The specific task examined in the study was the prediction of the functional similarity of two proteins from their sequences. It was chosen because the symmetric pair-input data needed for learning are readily available in biomedical databases: the sequences and functional annotations of hundreds of thousands of proteins are accessible in UniProt. The task is not particularly significant in the domain, however, because the task of directly classifying proteins by their functions is more straightforward and one of the major classification tasks in bioinformatics. The relationship between protein sequences and their functions is not trivial and hence low performance was expected in general. A symmetric feature representation of the protein pairs was constructed and K-nearest neighbour classifiers (Cover and Hart, 1967) with a range of K values were used as models.

The observed means and variances were as hypothesised, which encourages to initiate further studies. Experiments should be performed on other learning algorithms and data sets in order to get insights into the magnitudes of the effects. The prediction of protein–protein binding would be a particularly interesting task to explore not only because it is non-trivial to solve but also because negative instances are laborious to acquire. The absence of some object pairs from the data set can be expected to have an effect on the performance estimates. Cross-validation schemes intended for AB-type pairs or asymmetric pair-input data should also be considered as they are encountered in biomedical data.

This study considered only the identities of objects as the source of dependencies, but the approach to modify cross-validation schemes can be extended to other types of sources. Sequence similarity, functional similarity, or a shared membership of a protein family, for example, are properties that could be used to evaluate how a model would perform on newly discovered protein families. Sources that are external to the domain knowledge may also be utilised. In the task of protein–protein interaction extraction, dependencies emerge due to syntactic coordinations in which multiple interactions are expressed by the same words in a single sentence. The instances regarding these interactions share information due to the shared words and syntactic structures. The phrase

actin-binding proteins, profilin and cofilin

is an illustrative example of such dependencies. The *actin–profilin* and *actin–cofilin* pairs must be placed as a group either to the training set or to the validation set in order to maintain the independence of the two sets. The issue can be addressed by splitting the data at the document level (Airola et al., 2008). The approach discussed in Publication III has the advantage of considering dependencies across documents, but it does not take into account some dependencies within documents, which may occur due to shared topics or authorship (Pahikkala et al., 2006). The object-level approach to the partitioning of data sets should hence be explored using various knowledge-based methods of determining the existence of dependencies and compared to other approaches, such as sentence-level and document-level partitioning schemes, to discover which methods yield the most reliable estimates. Ideally, existing knowledge can be used to adjust cross-validation schemes to improve the reliability estimates of automatically inferred knowledge.

3.4 Ontology for patient education documents

The main purpose of patient education documents is to help patients to take responsibility of their health, which can be achieved through empowering patient education (Feste and Anderson, 1995). Ideally, patient education documents are readily available, understandable to patients (Vishnevetsky et al., 2018), and contain relevant high-quality information (Charnock et al., 1999; Vaartio-Rajalin et al., 2015).

Patient education documents and their content have been studied and modelled from the perspectives of understandability and patient's information needs (e.g., Johansson et al., 2004; Shoemaker et al., 2014; Kennedy et al., 2017; Brütting et al., 2018), focusing on the nursing science point of view. In contrast, patient education documents have been scarcely studied from the information technology point of view (e.g., Brown et al., 2001; Shapiro et al., 2005; Yang, 2005; Di Marco et al., 2006), and they are also in a marginal role in SNOMED CT and LOINC. Publication IV addresses the modelling of patient education documents from the knowledge representation perspective. The study is a step towards constructing a model of patient education documents and their readers to support the development of computational methods.

A patient education material ontology was developed to facilitate the use of information technology in the empowering patient education. The objectives of the ontology include the enhanced accessibility of documents through patient-friendly semantic queries, the consistency and relevance of content across documents in a large collection, and the personalisation of documents based on the characteristics of patients. To this end, the patient education material ontology models the use and content of patient education documents and the characteristics of their readers. In anticipation of the need for formal reasoning, the ontology was developed as an OWL2 ontology (Grau et al., 2008) because the OWL2 language is well-established and because there are many implementations of reasoners available for OWL2, such as HermiT (Glimm et al., 2014) and Pellet (Sirin et al., 2007).

In the patient education material ontology, the readers of documents can be described by their characteristics (such as gender or preferred language) but the documents can also be associated with these properties in order to describe their intended audience. The documents can further be characterised by their overall topics, such as a clinical action or a health condition, and by their uses in the care processes, expressed by the phase of care and the organisational unit they are used in. The ontology also includes classes for expressing the perspectives that are addressed in the text. A topic may be discussed, for example, from the perspectives of what patients should know or how they should act. The details of these perspectives and the actual topics discussed in patient education documents are not yet modelled in the ontology.

In addition to the design and implementation of the ontology structure, the ontology was populated with individuals that can be used to describe the intended uses of documents. Table 3.1 contains some examples of the classes of the ontology and of the individuals belonging to those classes. The individuals answer the questions what, when, and where regarding the overall purpose and topics of the documents. For example, a document titled "Home care after knee surgery" contains information on what patients should know when they are discharged from a hospital after a knee surgery. The document can be characterised by the individuals *knee* and *surgery*, which summarise the situation the document addresses, and by the individual *discharge*, which specifies the time when the document will be relevant for the patient. The ontology hence facilitates the organisation of documents in a user-friendly manner, such that the relevant documents can easily be accessed. This is an improvement over the solutions in which the

Table 3.1: *Examples of classes and individuals in the patient education material ontology.*

Class	Individuals
Phase of care	*discharge*, *preparation*
Body region topic	*chest*, *upper limb*
Body part topic	*knee*, *wrist*
Organ topic	*eye*, *heart*
Organ system topic	*nervous system*, *respiratory system*
Health condition topic	*inflammation*, *wound*
Clinical action topic	*physical therapy*, *surgery*
Symptom topic	*fever*, *pain*
Role	*patient*, *professional*
Language	*Finnish*, *Swedish*
Gender	*female*, *male*
Age group	*adult*, *child*

documents are only accessible via one property, such as medical specialisation or alphabetical order.

The Delphi method was originally developed to obtain knowledge on problems that are challenging to analyse due to the multitude of factors that would need to be taken into account. It consists of rounds during which the participants can express their opinions and give arguments for their views on a given topic. Its objective is to reach a consensus that incorporates the knowledge of domain experts into a solution that takes into account the variety of perspectives. (Powell, 2003; Keeney et al., 2006). It is commonly used in nursing science to leverage the domain knowledge of healthcare professionals (see, e.g., Trevelyan and Robinson, 2015; Foth et al., 2016). In this study, the Delphi method was used to discover the relevant entities to be included into the ontology. A Delphi panel was organised to obtain a set of entities that is suitable for describing patient education documents from the patients' perspective without losing the details and accuracy needed by the professionals. To this end, the opinions of domain experts on the relevance and clarity of the proposed entities and terminology were particularly considered.

The Delphi method was preferred over a data-driven approach because it can be used to assess the understandability of the proposed terms, which is essential in the construction of a patient-friendly ontology. It also makes possible to take into account the unwritten knowledge of domain experts that may not be evident in the available documents. However, since there is no clear, well-defined objective regarding which entities and relationships should be included in the ontology, existing medical resources were used as a reference. The main problem with these resources is their focus on the professionals' perspective: the resources have very exact entities that are needed to record care processes and patients' medical conditions but the details are unnecessary, or even harmful, in patient education. Hence, the resources were only used as a starting point in the development of the ontology. The proposed set of entities was refined to obtain a collection of entities that is compact, yet sufficient to describe patient education documents, and that contains terms that are understandable for both patients and professionals.

The descriptive statistics of the ontology and its construction process suggest that the ontology is suitable for its purpose. In particular, the ontology seems compact enough so that the terms needed by the users can easily be found. In its current state, the patient education material ontology is not yet finished, however. The Delphi method was an appropriate approach to collect the initial set of individuals but more work is needed to evaluate and refine the set. The ontology is currently in use in a local hospital district, and the feedback has revealed sections of the ontology that require further development. The individuals for describing the content of documents must be added before the ontology can be used for the analysis and manipulation of textual content. There is an ongoing research in this matter, aiming at the development of methods that can personalise patient education documents based on the characteristics of the patient and their care. The ontology has the framework for expressing rich relationships between entities but it is yet to be fully utilised.

3.5 Summarisation of electronic patient records

Professionals must familiarise themselves with patients, their conditions, and their prior care in order to provide best possible care, but the amount of information in electronic health records is overwhelming (Farri et al., 2012). This is particularly true for patients with long medical histories, which has motivated research efforts to develop and evaluate automatic summarisation methods (Pivovarov and Elhadad, 2015). The automatic summarisation of patient records has the potential to save professionals' time for actual care and improve the quality of care by providing information that may have been missed by the professionals.

Structured clinical data are an obvious target for visualisation and summarisation, and several studies have focused on the development of such methods (West et al., 2015). However, the summarisation of textual patient records is also vital because they contain information that is not available in structured form. Automatic text summarisation has been studied extensively in general (Nenkova and McKeown, 2012; Yao et al., 2017) but also specifically in the biomedical (Mishra et al., 2014; Nasr Azadani et al., 2018; Moradi, 2018) and healthcare domains (Sarker et al., 2012; Goldstein et al., 2017). Textual patient records pose specific challenges to summarisation, distinguishing them from the scientific literature and news articles (Pivovarov and Elhadad, 2015). Their prominent features include the patient's constantly evolving health condition, which may render some pieces of information obsolete, and the existence of specific topics that should be included to most summaries.

Publication V presents several methods to automatically summarise care episodes. They rely on word space models, which can be constructed in an unsupervised manner, to extract the most appropriate sentences from the source text. This approach facilitates the adaptation of the methods to the various settings that are present across hospitals and countries. To evaluate summarisation methods during their development, manual evaluation of generated summaries is a reliable approach, but it may not be feasible to use such a labour-intensive solution when a large number of candidate methods must be evaluated. Since summarisation methods based on word space models do not use text annotations to determine which pieces of knowledge should be included to summaries, the evaluation of their quality during their development should not rely on such information, either. Hence, automatic evaluation methods that can compare generated summaries to reference summaries are explored in Publication V.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) is a commonly applied set of evaluation methods in text summarisation (Lloret et al., 2018). It has been used to assess the quality of summaries in the biomedical and healthcare domains (e.g., Sarker et al., 2012; Nasr Azadani et al., 2018; Moradi, 2018). There are several ROUGE measures, all of which quantify text similarity by n-gram co-occurrences. The ROUGE variants differ by their details on how the n-gram co-occurrences are computed. ROUGE-N1 and ROUGE-N2, for example, use unigrams and bigrams, whereas ROUGE-L considers the longest common subsequences. In the study presented in Publication V, four ROUGE variants were used for automatic evaluation. They were also compared to manual evaluation performed by domain experts and found to correlate with the manual evaluation. ROUGE-N2 and ROUGE-L were found to be the best choices for automatic evaluation, with Spearman's ρ correlation coefficient 0.95. These results suggest that ROUGE measures can be used for automatic evaluation of care episode text summarisation methods.

The evaluation scheme used in this study mostly considered the presence of the specific pieces of information that are expected to be found in discharge summaries. While the ROUGE measures are useful in the initial development of text summarisation methods, the most promising methods should be subjected to extrinsic evaluation to assess their effectiveness in their intended purposes.

3.6 Large-scale sentiment analysis of news articles

Media have a role in shaping the public opinion on foreign countries because most individuals, lacking personal experience, must rely on media to obtain information about those countries (Perry, 1985; Entman, 2004). This applies to policy-makers, as well (O'Heffernan, 1991). Actors in societies can hence influence the public opinion and the policy making on specific issues through media (Soroka, 2003; Liu, 2006). As a result, countries are increasingly involved in managing their media images.

The officials of People's Republic of China (PRC) have the view that the Western media unjustly portraits PRC in a negative light, which has motivated PRC to engage herself to systematically improve her image in the Western world (Hu and Ji, 2012). The image of China in various media has been studied extensively, and both positive and negative images have been observed (e.g., Peng, 2004; Willnat and Luo, 2011; Xiang, 2013; Xie and Page, 2013). Text analysis in these studies mostly involved manual content analysis of small collections of text documents. Publication VI quantifies the sentiment in a large collection of Western news articles to discover whether there is an overall negative bias against China. As such, it complements the prior studies and provides further evidence to consider in the ongoing debate.

The Reuters Corpus, Volume 1 (RCV1) (Rose et al., 2002; Lewis et al., 2004) and Thomson Reuters Text Research Collection (TRC2) (Ounis et al., 2010) were chosen for the analysis. Reuters is a global news agency whose articles are widely used by news media in the Western world. It can hence be considered a representative sample of the Western mainstream media, but it does not cover alternative news outlets, particularly social media and other Internet news sources. The two corpora contain news coverage from August 1996 to August 1997 and from January 2008 to February 2009, respectively, which is convenient given that China adopted policies to improve her image in the Western countries between these time intervals. The analysis was therefore expected to reveal a positive trend in China's image over time. The other two hypothesis of the study were that 1) the China-related news articles are more negative than the articles about the East Asian regions that are ideologically closer to the Western world, and that 2) the articles about Chinese politics are more negative than those about Chinese culture because there are differences between the political structures of China and the Western countries and because China has focused on the promotion of her culture in the Western world.

The sentiment of a news article was quantified as the difference in the relative frequencies of the positive and negative words, which is the dictionary approach suggested by (Young and Soroka, 2012). The sentiment values of individual words were provided by the General Inquirer dictionary

(Stone et al., 1962), which is one of the most comprehensive dictionaries in the field (Young and Soroka, 2012). A simple dictionary-based classification scheme was used to classify the documents according to the regions (China, Japan, South Korea, Taiwan, Hong Kong) and themes (economy, politics, culture) they discuss. These pieces of information were needed to evaluate and compare the sentiments across regions and over time.

The area under the receiver operating characteristic curve (AUC) was applied to summarise the distributions of the document sentiments and was computed as the empirical AUC, equivalent to the Wilcoxon–Mann–Whitney U-statistic (Wilcoxon, 1945). Each target distribution was compared to the reference distribution that was computed from the other documents under the given conditions. As such, the observed AUC describes the relative sentiment of a set of documents against the other documents and can hence be considered as a baseline-adjusted sentiment. The DeLong test (DeLong et al., 1988) was then used to analyse whether there is a difference in the relative sentiments between two target sets of documents, which in this study differ by region, theme, or time period in a controlled manner. In short, the purpose of the AUC analysis was to isolate the phenomena of interest regardless of the characteristics of the sentiment distributions.

The AUC analysis concerns the comparison of sentiment distributions and requires document sentiments to be quantified with a continuous variable. Sentiment analysis can also be formulated as a time series analysis to discover trends or abrupt changes in sentiment over time (see, e.g., Young and Soroka, 2012). Such analyses use similar types of statistical variables as the AUC analysis. They provide more detailed information on trends than the AUC analysis but are limited to the analysis of time series data. Another approach to sentiment analysis is the sentiment classification task (see, e.g., Bakliwal et al., 2013; Chambers et al., 2015; Carrillo-de-Albornoz et al., 2018), in which documents are classified by their sentiment. It can be seen a task of obtaining sentiment distributions rather that of analysing them. In this task, sentiment lexicons may be used as sources of machine learning features. The sentiment classification and regression tasks can be used as preceding steps for further sentiment analysis, such as the AUC or time series analysis.

In this study, two types of analyses were performed with the overall sentiments and the sentiments within the given themes: China was compared to the other regions within each time period and the development of each region over time was analysed. For the China coverage, the culture theme was also compared to the other themes within each time period. The results provide evidence that China is not portrayed more negatively than the other East Asian regions. On the contrary, the news articles about China tend to be more positive overall than those about the other regions. A positive trend over time was also observed for China. In the latter time period, the news articles about China's economy and culture were more positive than those about China's politics, which is not surprising given the differences between the political systems of China and the Western countries. Unfortunately, due to the issues with the document classification, the results regarding the culture sentiment analysis should be treated with caution.

In general, the proposed large-scale AUC analysis can be used when sets of documents should be compared but the sentiment baseline is expected to vary between the sets. The reliability of the analysis depends on the quality of the document classification and the appropriateness of the selected sentiment quantification method. This study could be improved by applying supervised machine learning to the classification of documents and by developing a more detailed sentiment quantification method.

The use of machine learning techniques to classify the articles by their regions or themes was beyond the possibilities of this study because of limited human resources. The quality of the classification could most likely have been significantly improved by training a classification model on the well-annotated news articles of RCV1. This approach was not taken because there was the risk that the classification error rate on RCV1 would have been lower than on TRC2 due to the differences in the characteristics of the corpora, which would have interfered with the AUC analysis. Unlike RCV1, TRC2 seems to contain raw text streams rather than well-prepared and annotated articles. There may also be significant differences between the corpora in the content and styles of the articles due to how news reporting has changed over the years. Training a classification model on TRC2 would have required notable resources for manual annotation.

For a more detailed sentiment analysis, an appropriate approach would be to take advantage of syntactic parsing (Di Caro and Grella, 2013) and information extraction. This would allow to analyse which specific entities or topics are positively or negatively discussed and would hence provide more accurate results. However, such an analysis would need to focus on a small number of themes because information extraction on all possible topics is intractable. Another possible direction for future research efforts is to apply the AUC analysis in the healthcare domain, for example, to analyse patients' cancer treatment experiences across cancer types or over time.

Chapter 4

Summary

Knowledge representation and text mining methods are valuable tools in the analysis and management of text masses and domain knowledge. Knowledge representations are useful in organising, exchanging, and analysing domain knowledge, whereas text mining methods can be used to extract knowledge from textual data sources. This thesis explores three domains and their characteristics in relation to knowledge representation and text mining. In general, methods can be used across domains but the domain-specific needs of the applications must be addressed by adapting the methods.

In the biomedical domain, knowledge may be derived from the physical and chemical properties of molecules, which results in knowledge that can be expressed with symbolic or mathematical representations. Biomedical scientific articles contain a minimal amount of ambiguity because their content is the result of careful analyses and because the reported interpretations of the data are as objective as possible. The knowledge is expressed in an exact manner with a limited vocabulary.

The healthcare domain shares the objective of unbiased and exact reporting with the biomedical domain. However, the biological processes involved in health conditions and the documentative nature of patient records lead to textual data that is incomplete and has notable variation in the degree of interpretations made by the professionals. Hence, the reader of the text has an elevated responsibility of making further interpretations from the reported information and the computational methods applied to the text must be able to handle uncertainties in the input. Patient education documents share topics with patient records but are written for patients, which must be taken into account in the applied methodology. The knowledge representations in the field tend to focus on the documentation and statistical analysis of healthcare processes and the status of individual patients, although there are also resources that capture universal knowledge regarding the domain.

In political science, scientific studies analyse textual data that is not necessarily objective. Texts that have a political agenda attempt to influence the audience, which should be addressed in the analysis of the text. Even news paper articles may be, intentionally or unintentionally, biased, although journalistic best practices emphasise the reliability and objectivity of reporting. Consequently, the texts contain the interpretations of the reporters about the occurred events, and such interpretations may vary greatly between the involved parties. The analysis of media and political texts must therefore take into account the possible deviation of the textual content from the truth. Regardless, knowledge representations in political science are similar to those in other fields.

The aims of this thesis concern the analysis of the scientific literature to extract biomedical knowledge, the development of computational methods and resources to improve the quality of health care, and the large-scale sentiment analysis of news media. The six original publications of this thesis also illustrate how the objectives of the study affect the selection of analysis methods and how the methods may need to be adapted to address the requirements arising from the characteristics of the domain.

Publications I and II discuss the extraction of biomolecular events from scientific articles and the transformation of such knowledge into form that is more convenient for domain experts to use. They argue that the extraction of complex relationships from text is a viable approach and that complex relationships can be reliably transformed into binary relationships, which are commonly used by the domain experts. These results contribute to the information extraction methodology in biomedical text mining by facilitating new ways in which knowledge can be obtained from text to be used in biomedical applications. Publication III argues for the use of a specific cross-validation method for pair-input data that takes into account the dependencies that emerge due to shared pair members. Its main idea can also be adapted to other types of knowledge and to the text mining of protein–protein interactions. The results of this study help to more effectively utilise biomedical knowledge in analyses that rely on machine learning.

Publication IV describes the patient education material ontology that was developed in response to the lack of a patient-friendly knowledge resource that could be used to improve the accessibility and content quality of patient education documents. It argues that the ontology should be designed from the perspective of patients, a property that the existing resources lack in the domain. The ontology serves as a basis for methods and applications intended improve the care given to patients, such as the personalisation of patient education documents. It is a contribution towards the use of knowledge representations in patient education. Publication V explores the automatic summarisation of patient records. It does not use knowledge representations in order to remain content-independent. One of the contributions of this study to the text mining of patient records is the conclusion that the well-known ROUGE measures can be used in the automatic evaluation of summarisation methods, which helps to accelerate their development. The availability of high-quality summarisation methods can ease the workload of the healthcare professionals.

Publication VI analyses the image of China in the Western media through sentiment analysis. It quantifies the tendency of news articles to exhibit positive or negative tone and argues that there is no overall systematic bias against China. The study adapts methods that are used in the biomedical domain and hence represents an example of the possibilities of methodological adaptations. Its methodological contribution is the demonstration how large-scale corpora of news articles can be analysed with AUC to reveal phenomena that are commonly studied using labour-intensive methods in the field of political science.

There are promising future directions in each of the studies presented in this thesis. Of particular interest is the investigation of pair-input cross-validation with other biomedical data sets, the use of the patient education material ontology in the personalisation of patient education documents, and the further analysis of social and news media to discover trends in sentiment. These studies provide ample opportunities to continue knowledge representation and text mining research in the domains of biomedicine, healthcare, and political science.

Bibliography

Abercrombie, G. and Batista-Navarro, R. T. (2018). Identifying opinion-topics and polarity of parliamentary debate motions. In Balahur, A., Mohammad, S. M., Hoste, V. and Klinger, R. (eds.), *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 280–285). Stroudsburg, PA, USA: Association for Computational Linguistics.

Aggarwal, C. C. and Zhai, C. (eds.) (2012). *Mining Text Data*. Boston, MA, USA: Springer.

Ahmad, K., Daly, N. and Liston, V. (2011). What is new? News media, general elections, sentiment, and named entities. In Bandyopadhyay, S. and Okumurra, M. (eds.), *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology* (pp. 80–88). Asian Federation of Natural Language Processing.

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F. and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9 Suppl 11:S2.

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R. and Wang, X. (2008). Assisted curation: does text mining really help? *Pacific Symposium on Biocomputing* 13:556–567.

Ananiadou, S. and McNaught, J. (eds.) (2006). *Text Mining for Biology and Biomedicine*. Boston, MA, USA: Artech House.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* 4:40–79.

Aukia, J., Heimonen, J., Pahikkala, T. and Salakoski, T. (2017). Automated quantification of Reuters news using a receiver operating characteristic curve analysis: the Western media image of China. *Global Media and China* 2:251–268.

Bakliwal, A., Foster, J., Puil, J. van der, O'Brien, R., Tounsi, L. and Hughes, M. (2013). Sentiment analysis of political tweets: towards an accurate classifier. In Danescu-Niculescu-Mizil, C., Farzindar, A., Gamon, M., Inkpen, D. and Nagarajan, M. (eds.), *Proceedings of the Workshop on*

Language Analysis in Social Media (pp. 49–58). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bekoulis, G., Deleu, J., Demeester, T. and Develder, C. (2018). Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications* 114:34–45.

Bhatia, S. and P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. In Balahur, A., Mohammad, S. M., Hoste, V. and Klinger, R. (eds.), *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 79–84). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bhise, V., Rajan, S. S., Sittig, D. F., Vaghani, V., Morgan, R. O., Khanna, A. and Singh, H. (2018). Electronic health record reviews to measure diagnostic uncertainty in primary care. *Journal of Evaluation in Clinical Practice* 24:545–551.

Björne, J., Ginter, F., Heimonen, J., Pyysalo, S. and Salakoski, T. (2009). Learning to extract biological event and relation graphs. In Jokinen, K. and Bick, E. (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics* (pp. 18–25). Northern European Association for Language Technology.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T. and Salakoski, T. (2011). Extracting contextualized complex biological events with rich graph-based feature sets. *Computational Intelligence* 27:541–557.

Björne, J., Pyysalo, S., Ginter, F. and Salakoski, T. (2008). How complex are complex protein-protein interactions? In Salakoski, T., Rebholz-Schuhmann, D. and Pyysalo, S. (eds.), *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (pp. 125–128). Turku, Finland: Turku Centre for Computer Science.

Björne, J. and Salakoski, T. (2011). Generalizing biomedical event extraction. In Tsujii, J., Kim, J.-D. and Pyysalo, S. (eds.), *Proceedings of BioNLP Shared Task 2011 Workshop* (pp. 183–191). Stroudsburg, PA, USA: Association for Computational Linguistics.

Björne, J. and Salakoski, T. (2013). TEES 2.1: automated annotation scheme learning in the BioNLP 2013 Shared Task. In Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-j., Ohta, T., Pyysalo, S. and Zweigenbaum, P. (eds.), *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 16–25). Stroudsburg, PA, USA: Association for Computational Linguistics.

Björne, J. and Salakoski, T. (2018). Biomedical event extraction using convolutional neural networks and dependency parsing. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S. and Tsujii, J. (eds.), *Proceedings of the BioNLP 2018 workshop* (pp. 98–108). Stroudsburg, PA, USA: Association for Computational Linguistics.

Boyd, A. D., Dunn Lopez, K., Lugaresi, C., Macieira, T., Sousa, V., Acharya, S., Balasubramanian, A., Roussi, K., Keenan, G. M., Lussier, Y. A., Li, J. J., Burton, M. and Di Eugenio, B. (2018). Physician nurse care: a new use of UMLS to measure professional contribution: are we talking about the same patient a new graph matching algorithm? *International Journal of Medical Informatics* 113:63–71.

Brachman, R. and Levesque, H. (2004). *Knowledge Representation and Reasoning.* San Francisco, CA, USA: Morgan Kaufmann Publishers.

Brisson, C. (2003). Plurals, all, and the nonuniformity of collective predication. *Linguistics and Philosophy* 26:129–184.

Brown, S. H., Lincoln, M., Hardenbrook, S., Petukhova, O. N., Rosenbloom, S. T., Carpenter, P. and Elkin, P. (2001). Derivation and evaluation of a document-naming nomenclature. *Journal of the American Medical Informatics Association* 8:379–390.

Brown, S., Bonial, C., Obrst, L. and Palmer, M. (2017). The Rich Event Ontology. In Caselli, T., Miller, B., Erp, M. van, Vossen, P., Palmer, M., Hovy, E., Mitamura, T. and Caswell, D. (eds.), *Proceedings of the Events and Stories in the News Workshop* (pp. 87–97). Stroudsburg, PA, USA: Association for Computational Linguistics.

Brütting, J., Reinhardt, L., Bergmann, M., Schadendorf, D., Weber, C., Tilgen, W., Berking, C. and Meier, F. (2018). Quality, readability, and understandability of German booklets addressing melanoma patients. *Journal of cancer education.* DOI: 10.1007/s13187-018-1369-x.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K. and Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine* 33:139–155.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E. and ENVO Consortium (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 4:43.

Carrillo-de-Albornoz, J., Rodríguez Vidal, J. and Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *PLoS ONE* 13:e0207996.

Chambers, N., Bowen, V., Genco, E., Tian, X., Young, E., Harihara, G. and Yang, E. (2015). Identifying political sentiment between nation states with social media. In Màrquez, L., Callison-Burch, C. and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 65–75). Stroudsburg, PA, USA: Association for Computational Linguistics.

Charnock, D., Shepperd, S., Needham, G. and Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology and Community Health* 53:105–111.

Chen, H., Fuller, S. S., Friedman, C. and Hersh, W. (eds.) (2005). *Medical Informatics.* Boston, MA, USA: Springer.

Cios, K. J., Swiniarski, R. W., Pedrycz, W. and Kurgan, L. A. (2007). *Data Mining – A Knowledge Discovery Approach.* Boston, MA, USA: Springer.

Clark, A., Fox, C. and Lappin, S. (eds.) (2013). *The Handbook of Computational Linguistics and Natural Language Processing.* West Sussex, United Kingdom: Wiley-Blackwell.

Cohen, K. B. and Demner-Fushman, D. (2014). *Biomedical Natural Language Processing.* Amsterdam, Netherlands: John Benjamins Publishing Company.

Considine, J., Trotter, C. and Currey, J. (2016). Nurses' documentation of physiological observations in three acute care settings. *Journal of Clinical Nursing* 25:134–143.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13:21–27.

Cyganiak, R., Wood, D. and Lanthaler, M. (eds.) (2014). *RDF 1.1 Concepts and Abstract Syntax.* W3C Recommendation 25 February 2014. URL: `https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/`.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845.

Di Caro, L. and Grella, M. (2013). Sentiment analysis via dependency parsing. *Computer Standards & Interfaces* 35:442–453.

Di Marco, C., Bray, P., Covvey, H. D., Cowan, D. D., Di Ciccio, V., Hovy, E., Lipa, J. and Yang, C. (2006). Authoring and generation of individualized patient education materials. *AMIA Annual Symposium Proceedings* 2006:195–199.

Diermeier, D., Godbout, J.-F., Yu, B. and Kaufmann, S. (2011). Language and ideology in congress. *British Journal of Political Science* 42:31–55.

Dijk, J. F. M. van, Vervoort, S. C. J. M., Wijck, A. J. M. van, Kalkman, C. J. and Schuurmans, M. J. (2016). Postoperative patients' perspectives on rating pain: a qualitative study. *International Journal of Nursing Studies* 53:260–269.

Driscoll, K. and Thorson, K. (2015). Searching and clustering methodologies: connecting political communication content across platforms. *The ANNALS of the American Academy of Political and Social Science* 659:134–148.

Entman, R. M. (2004). *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy.* Chicago, IL, USA: University of Chicago Press.

Entman, R. M. (2007). Framing bias: media in the distribution of power. *Journal of Communication* 57:163–173.

Evans, R. S. (2016). Electronic health records: then, now, and in the future. *Yearbook of Medical Informatics* Suppl 1:48–61.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. and D'Eustachio, P. (2018). The Reactome pathway knowledgebase. *Nucleic Acids Research* 46:D649–D655.

Farri, O., Pieckiewicz, D. S., Rahman, A. S., Adam, T. J., Pakhomov, S. V. and Melton, G. B. (2012). A qualitative analysis of EHR clinical document synthesis by clinicians. *AMIA Annual Symposium Proceedings* 2012:1211–1220.

Feste, C. and Anderson, R. M. (1995). Empowerment: from philosophy to practice. *Patient Education and Counseling* 26:139–144.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D. and Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In Riloff, E., Chiang, D., Julia, H. and Jun'ichi, T. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3570–3580). Stroudsburg, PA, USA: Association for Computational Linguistics.

Flasiński, M. (2016). *Introduction to Artificial Intelligence.* Cham, Switzerland: Springer.

Fortuny, E. J. de, De Smedt, T., Martens, D. and Daelemans, W. (2012). Media coverage in times of political crisis: a text mining approach. *Expert Systems with Applications* 39:11616–11622.

Foth, T., Efstathiou, N., Vanderspank-Wright, B., Ufholz, L.-A., Dütthorn, N., Zimansky, M. and Humphrey-Murto, S. (2016). The use of Delphi and Nominal Group Technique in nursing education: a review. *International Journal of Nursing Studies* 60:112–120.

Friedman, C., Kra, P. and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35:222–235.

GDELT Project (2015). *The GDELT Global Knowledge Graph (GKG) Data Format Codebook v2.1.* URL: `http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf`.

Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* 45:D331–D338.

Gene Ontology Consortium, Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin,

G. M. and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:25–29.

Gerner, D. J., Abu-Jabr, R., Schrodt, P. A. and Yilmaz, Ö. (2002). *Conflict and Mediation Event Observations (CAMEO): a new event data framework for the analysis of foreign policy interactions.* Paper prepared for delivery at the Annual Meeting of the International Studies Association, New Orleans, March 2002. URL: https://pdfs.semanticscholar.org/775d/7f7262ffb42972e5b87a245bc4b63c20396d.pdf.

Giordano, J., Abramson, K. and Boswell, M. V. (2010). Pain assessment: subjectivity, objectivity, and the use of neurotechnology. *Pain Physician* 13:305–315.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G. and Wang, Z. (2014). HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning* 53:245–269.

Goldstein, A., Shahar, Y., Orenbuch, E. and Cohen, M. J. (2017). Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. *Artificial Intelligence in Medicine* 82:20–33.

Gonçalves, P. D. B., Sequeira, C. A. C. and Silva, M. A. T. C. P. e (2019). Content analysis of nursing diagnoses in mental health records in Portugal. *International Nursing Review* 66:199–208.

González, C. R. and Abu-Mostafa, Y. S. (2015). Mismatched training and test distributions can outperform matched ones. *Neural Computation* 27:365–387.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P. and Sattler, U. (2008). OWL 2: the next step for OWL. *Journal of Web Semantics* 6:309–322.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H. and Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* 46:D380–D386.

Häyrinen, K., Saranto, K. and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics* 77:291–304.

Heimonen, J., Björne, J. and Salakoski, T. (2010). Reconstruction of semantic relationships from their projections in biomolecular domain. In Cohen, K. B., Demner-Fushman, D., Ananiadou, S., Pestian, J., Tsujii, J. and Webber, B. (eds.), *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (pp. 108–116). Stroudsburg, PA, USA: Association for Computational Linguistics.

Heimonen, J., Danielsson-Ojala, R., Salakoski, T., Lundgrén-Laine, H. and Salanterä, S. (2018). Ontology development for patient education documents using a professional- and patient-oriented Delphi method. *CIN: Computers, Informatics, Nursing* 36:448–457.

Heimonen, J., Pyysalo, S., Ginter, F. and Salakoski, T. (2008). Complex-to-pairwise mapping of biological relationships using a semantic network representation. In Salakoski, T., Rebholz-Schuhmann, D. and Pyysalo, S. (eds.), *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (pp. 45–52). Turku, Finland: Turku Centre for Computer Science.

Heimonen, J., Salakoski, T. and Pahikkala, T. (2014). Properties of object-level cross-validation schemes for symmetric pair-input data. In Fränti, P., Brown, G., Loog, M., Escolano, F. and Pelillo, M. (eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 384–393). Berlin Heidelberg, Germany: Springer.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V. and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* 43:D512–D520.

Hu, Z. and Ji, D. (2012). Ambiguities in communicating with the world: the "going-out" policy of China's media and its multilayered contexts. *Chinese Journal of Communication* 5:32–37.

Jakus, G., Milutinović, V., Omerović, S. and Tomažič, S. (2013). *Concepts, Ontologies, and Knowledge Representation.* New York, NY, USA: Springer.

Johansson, K., Salanterä, S., Katajisto, J. and Leino-Kilpi, H. (2004). Written orthopedic patient education materials from the point of view of empowerment by education. *Patient Education and Counseling* 52:175–181.

Kaal, B., Maks, I. and Elfrinkhof, A. van (eds.) (2014). *From Text to Political Positions.* Amsterdam, Netherlands: John Benjamins Publishing Company.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research* 47:D590–D595.

Karcioglu, O., Topacoglu, H., Dikme, O. and Dikme, O. (2018). A systematic review of the pain scales in adults: which to use? *American Journal of Emergency Medicine* 36:707–714.

Kauhanen, L., Murtola, L.-M., Heimonen, J., Leskinen, T., Kalliokoski, K., Raivo, E., Salakoski, T. and Salanterä, S. (2014). Documentation of the clinical phase of the cardiac rehabilitation process in a Finnish university hospital district. In Saranto, K., Castrén, M., Kuusela, T., Hyrynsalmi, S. and Ojala, S. (eds.), *Safe and Secure Cities* (pp. 57–67). Cham, Switzerland: Springer.

Keeney, S., Hasson, F. and McKenna, H. (2006). Consulting the oracle: ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing* 53:205–212.

Kennedy, D., Wainwright, A., Pereira, L., Robarts, S., Dickson, P., Christian, J. and Webster, F. (2017). A qualitative study of patient education needs for hip and knee replacement. *BMC Musculoskeletal Disorders* 18:413.

Kikugawa, S., Nishikata, K., Murakami, K., Sato, Y., Suzuki, M., Altaf-Ul-Amin, M., Kanaya, S. and Imanishi, T. (2012). PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset. *BMC Systems Biology* 6 Suppl 2:S7.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. (2009). Overview of BioNLP'09 Shared Task on event extraction. In Tsujii, J. (ed.), *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task* (pp. 1–9). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kim, J.-D., Ohta, T. and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10.

Kim, J.-D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N. and Tsujii, J. (2011). Overview of BioNLP Shared Task 2011. In Tsujii, J., Kim, J.-D. and Pyysalo, S. (eds.), *Proceedings of BioNLP Shared Task 2011 Workshop* (pp. 1–6). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kim, J.-D., Wang, Y., Colic, N., Beak, S. H., Kim, Y. H. and Song, M. (2016). Refactoring the Genia event extraction shared task toward a general framework for IE-driven KB development. In Nédellec, C., Bossy, R. and Kim, J.-D. (eds.), *Proceedings of the 4th BioNLP Shared Task Workshop* (pp. 23–31). Stroudsburg, PA, USA: Association for Computational Linguistics.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Laulederkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Lourghi, H., Della Rocca, M. G., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C. and Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* 47:D1018–D1027.

Komulainen, J. (ed.) (2011). *Tautiluokitus ICD-10*, 3rd ed. Helsinki, Finland: National Institute for Health and Welfare.

Kvist, M. and Savolainen, T. (eds.) (2010). *ICPC-2 Perusterveydenhuollon kansainvälinen luokitus*. Helsinki, Finland: The Association of Finnish Local and Regional Authorities.

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

Leetaru, K. and Schrodt, P. A. (2013). *GDELT: global data on events, location and tone, 1979-2012*. Paper presented at the International Studies Association meetings, San Francisco, April 2013. URL: `http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf`.

Lehtonen, J., Lehtovirta, J. and Mäkelä-Bengs, P. (2013). *THL-Toimenpideluokitus*. Helsinki, Finland: National Institute for Health and Welfare.

Lenert, L. A. (2016). Toward medical documentation that enhances situational awareness learning. *AMIA Annual Symposium Proceedings* 2016:763–771.

Lewis, D. D., Yang, Y., Rose, T. G. and Li, F. (2004). RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5:361–397.

Liljamo, P., Kinnunen, U.-M. and Ensio, A. (2012). *FinCC-luokituskokonaisuuden käyttöopas - SHTaL 3.0, SHToL 3.0, SHTuL 1.0*. Helsinki, Finland: National Institute for Health and Welfare.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In Moens, M.-F. and Szpakowicz, S. (eds.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, Y., Mehta, S., Küçük-McGinty, H., Turner, J. P., Vidovic, D., Forlin, M., Koleti, A., Nguyen, D.-T., Jensen, L. J., Guha, R., Mathias, S. L., Ursu, O., Stathias, V., Duan, J., Nabizadeh, N., Chung, C., Mader, C., Visser, U., Yang, J. J., Bologa, C. G., Oprea, T. I. and Schürer, S. C. (2017). Drug target ontology to classify and integrate drug discovery data. *Journal of Biomedical Semantics* 8:50.

Liu, X. (2006). *Modeling Bilateral International Relations: The Case of U.S.-China Interactions*. New York, NY, USA: Palgrave Macmillan.

Lloret, E., Plaza, L. and Aker, A. (2018). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation* 52:101–148.

Lofi, C., Nieke, C. and Collier, N. (2014). Discriminating rhetorical analogies in social media. In Wintner, S., Goldwater, S. and Riezler, S. (eds.), *Proceedings of the 14th Conference of the European Chapter of the Asso-*

ciation for Computational Linguistics (pp. 560–568). Stroudsburg, PA, USA: Association for Computational Linguistics.

Marneffe, M.-C. de and Manning, C. D. (2008a). *Stanford typed dependencies manual*. Revised for the Stanford Parser v. 3.7.0 in September 2016. URL: `https://nlp.stanford.edu/software/dependencies_manual.pdf`.

Marneffe, M.-C. de and Manning, C. D. (2008b). The Stanford typed dependencies representation. In Bos, J., Briscoe, E., Cahill, A., Carroll, J., Clark, S., Copestake, A., Flickinger, D., Genabith, J. van, Hockenmaier, J., Joshi, A., Kaplan, R., King, T. H., Kuebler, S., Lin, D., Lønning, J. T., Manning, C., Miyao, Y., Nivre, J., Oepen, S., Sagae, K., Xue, N. and Zhang, Y. (eds.), *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation* (pp. 1–8). Coling 2008 Organizing Committee.

McAuley, J. W. (2003). *An Introduction to Politics, State and Society*. London, United Kingdom: SAGE Publications.

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., Williams, W., Case, J. and Maloney, P. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical Chemistry* 49:624–633.

Meldal, B. H., Forner-Martinez, O., Costanzo, M. C., Dana, J., Demeter, J., Dumousseau, M., Dwight, S. S., Gaulton, A., Licata, L., Melidoni, A. N., Ricard-Blum, S., Roechert, B., Skyzypek, M. S., Tiwari, M., Velankar, S., Wong, E. D., Hermjakob, H. and Orchard, S. (2015). The complex portal – an encyclopaedia of macromolecular complexes. *Nucleic Acids Research* 43:D479–D484.

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J. and Del Fiol, G. (2014). Text summarization in the biomedical domain: a systematic review of recent research. *Journal of Biomedical Informatics* 52:457–467.

Moen, H., Peltonen, L.-M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T. and Salanterä, S. (2016). Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine* 67:25–37.

Moradi, M. (2018). CIBS: a biomedical text summarizer using topic-based sentence clustering. *Journal of Biomedical Informatics* 88:53–61.

Nasr Azadani, M., Ghadiri, N. and Davoodijam, E. (2018). Graph-based biomedical text summarization: an itemset mining and sentence clustering approach. *Journal of Biomedical Informatics* 84:42–58.

Nédellec, C. (2005). Learning Language in Logic - genic interaction extraction challenge. In Cussens, J. and Nédellec, C. (eds.), *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning* (pp. 31–37).

Nédellec, C., Bossy, R. and Kim, J.-D. (eds.) (2016). *Proceedings of the 4th BioNLP Shared Task Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-j., Ohta, T., Pyysalo, S. and Zweigenbaum, P. (2013). Overview of BioNLP Shared Task 2013. In Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-j., Ohta, T., Pyysalo, S. and Zweigenbaum, P. (eds.), *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 1–7). Stroudsburg, PA, USA: Association for Computational Linguistics.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In Aggarwal, C. C. and Zhai, C. (eds.), *Mining Text Data* (pp. 43–76). Boston, MA, USA: Springer.

Numan, T., Boogaard, M. van den, Kamper, A. M., Rood, P. J. T., Peelen, L. M. and Slooter, A. J. C. (2017). Recognition of delirium in post-operative elderly patients: a multicenter study. *Journal of the American Geriatrics Society* 65:1932–1938.

O'Heffernan, P. (1991). *Mass media and American foreign policy: Insider perspectives on global journalism and the foreign policy process*. Norwood, NJ, USA: Ablex Publishing.

Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* 17:155–161.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Roey, K. van, Cesareni, G. and Hermjakob, H. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* 42:D358–D363.

Ounis, I., Soboroff, I. M. and Macdonald, C. (2010). Overview of the TREC-2010 Blog track. In Voorhees, E. M. and Buckland, L. P. (eds.), *Proceedings of the Nineteenth Text REtrieval Conference*. USA: National Institute of Standards and Technology.

Pahikkala, T., Boberg, J. and Salakoski, T. (2006). Fast n-fold cross-validation for regularized least-squares. In Honkela, T., Raiko, T., Kortela, J. and Harri, V. (eds.), *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence* (pp. 83–90). Espoo, Finland: Finnish Artificial Intelligence Society.

Pahikkala, T., Suominen, H. and Boberg, J. (2012). Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Machine Learning* 87:381–407.

Park, Y. and Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nature Methods* 9:1134–1136.

Peng, Z. (2004). Representation of China: an across time analysis of coverage in the New York Times and Los Angeles Times. *Asian Journal of Communication* 14:53–67.

Pennings, P., Keman, H. and Kleinnijenhuis, J. (2006). *Doing Research in Political Science*, 2nd ed. London, United Kingdom: SAGE Publications.

Perry, D. K. (1985). The mass media and inference about other nations. *Communication Research* 12:595–614.

Pivovarov, R. and Elhadad, N. (2015). Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association* 22:938–947.

Portin, P. and Wilkins, A. (2017). The evolving definition of the term "gene". *Genetics* 205:1353–1364.

Powell, C. (2003). The Delphi technique: myths and realities. *Journal of Advanced Nursing* 41:376–382.

Praag, P. van (ed.) (2017). *Political Science and Changing Politics.* Amsterdam, Netherlands: Amsterdam University Press.

Pudas-Tähkä, S.-M., Axelin, A., Aantaa, R., Lund, V. and Salanterä, S. (2014). Translation and cultural adaptation of an objective pain assessment tool for Finnish ICU patients. *Scandinavian Journal of Caring Sciences* 28:885–894.

Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F. and Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics* 9 Suppl 3:S6.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J. and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8:50.

Reddy, S., Täckström, O., Petrov, S., Steedman, M. and Lapata, M. (2017). Universal semantic parsing. In Palmer, M., Hwa, R. and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 89–101). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rodrigues, J.-M., Robinson, D., Della Mea, V., Campbell, J., Rector, A., Schulz, S., Brear, H., Üstün, B., Spackman, K., Chute, C. G., Millar, J., Solbrig, H. and Brand Persson, K. (2015). Semantic alignment between ICD-11 and SNOMED CT. *Studies in Health Technology and Informatics* 216:790–794.

Rose, T., Stevenson, M. and Whitehead, M. (2002). The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In Rodríguez, M. G. and Suarez Araujo, C. P. (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 827–833). Paris, France: European Language Resources Association.

Rosse, C. and Mejino Jr, J. L. V. (2008). The Foundational Model of Anatomy ontology. In Burger, A., Davidson, D. and Baldock, R. (eds.), *Anatomy Ontologies for Bioinformatics: Principles and Practice* (pp. 59–117). London, United Kingdom: Springer.

Rubin, V., Conroy, N., Chen, Y. and Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In Fornaciari, T., Fitzpatrick, E. and Bachenko, J. (eds.), *Proceedings of the Second Workshop on Computational Approaches to Deception Detection* (pp. 7–17). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S. and Sedlmair, M. (2018). More than bags of words: sentiment analysis with word embeddings. *Communication Methods and Measures* 12:140–157.

El-Sappagh, S., Franda, F., Ali, F. and Kwak, K.-S. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making* 18:76.

Sarker, A., Mollá, D. and Paris, C. (2012). Extractive summarisation of medical documents using domain knowledge and corpus statistics. *Australasian Medical Journal* 5:478–481.

Sarntivijai, S., Lin, Y., Xiang, Z., Meehan, T. F., Diehl, A. D., Vempati, U. D., Schürer, S. C., Pang, C., Malone, J., Parkinson, H., Liu, Y., Takatsuki, T., Saijo, K., Masuya, H., Nakamura, Y., Brush, M. H., Haendel, M. A., Zheng, J., Stoeckert, C. J., Peters, B., Mungall, C. J., Carey, T. E., States, D. J., Athey, B. D. and He, Y. (2014). CLO: the cell line ontology. *Journal of Biomedical Semantics* 5:37.

Schulz, S. and Martinez-Costa, C. (2015). Harmonizing SNOMED CT with BioTopLite: an exercise in principled ontology alignment. *Studies in Health Technology and Informatics* 216:832–836.

Segers, R., Caselli, T. and Vossen, P. (2017). The Circumstantial Event Ontology (CEO). In Caselli, T., Miller, B., Erp, M. van, Vossen, P., Palmer, M., Hovy, E., Mitamura, T. and Caswell, D. (eds.), *Proceedings of the Events and Stories in the News Workshop* (pp. 37–41). Stroudsburg, PA, USA: Association for Computational Linguistics.

Shapiro, J. S., Bakken, S., Hyun, S., Melton, G. B., Schlegel, C. and Johnson, S. B. (2005). Document ontology: supporting narrative documents in electronic health records. *AMIA Annual Symposium Proceedings* 2005:684–688.

Shoemaker, S. J., Wolf, M. S. and Brach, C. (2014). Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Education and Counseling* 96:395–403.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A. and Katz, Y. (2007). Pellet: a practical OWL-DL reasoner. *Journal of Web Semantics* 5:51–53.

Sjöblom, O., Murtola, L.-M., Heimonen, J., Kauhanen, L., Laippala, V., Lundgrén-Laine, H., Salakoski, T. and Salanterä, S. (2013). Using cluster analysis to identify weak signals of lethal trends in aviation and healthcare documentation. *International Journal of Networking and Virtual Organisations* 13:66–80.

Soroka, S. N. (2003). Media, public opinion, and foreign policy. *Harvard International Journal of Press/Politics* 8:27–48.

Stone, P. J., Bales, R. F., Namenwirth, J. Z. and Ogilvie, D. M. (1962). The General Inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Computers in Behavioral Science* 7:484–498.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J. and Mering, C. von (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* 45:D362–D368.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. *Pacific Symposium on Biocomputing* 5:541–552.

Tian, L., Lai, C. and Moore, J. (2018). Polarity and intensity: the two aspects of sentiment analysis. In Zadeh, A., Liang, P. P., Morency, L.-P., Poria, S., Cambria, E. and Scherer, S. (eds.), *Proceedings of Grand Challenge and Workshop on Human Multimodal Language* (pp. 40–47). Stroudsburg, PA, USA: Association for Computational Linguistics.

Trevelyan, E. G. and Robinson, N. (2015). Delphi methodology in health research: how to do it? *European Journal of Integrative Medicine* 7:423–428.

UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45:D158–D169.

Uronen, L., Heimonen, J., Puukka, P., Martimo, K.-P., Hartiala, J. and Salanterä, S. (2017). Health check documentation of psychosocial factors using the WAI. *Occupational Medicine* 67:151–154.

Vaartio-Rajalin, H., Huumonen, T., Iire, L., Jekunen, A., Leino-Kilpi, H., Minn, H. and Paloniemi, J. (2015). Patient education process in oncologic context: what, why, and by whom? *Nursing research* 64:381–390.

Veres, D. V., Gyurkó, D. M., Thaler, B., Szalay, K. Z., Fazekas, D., Korcsmáros, T. and Csermely, P. (2015). ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Research* 43:D485–493.

Vishnevetsky, J., Walters, C. B. and Tan, K. S. (2018). Interrater reliability of the Patient Education Materials Assessment Tool (PEMAT). *Patient Education and Counseling* 101:490–496.

Wang, A. Y., Sable, J. H. and Spackman, K. A. (2002). The SNOMED clinical terms development process: refinement and analysis of content. *Proceedings of the AMIA Symposium* 2002:845–849.

West, V. L., Borland, D. and Hammond, W. E. (2015). Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association* 22:330–339.

Wielemaker, J., Schrijvers, T., Triska, M. and Lager, T. (2012). SWI-Prolog. *Theory and Practice of Logic Programming* 12:67–96.

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

Wilkerson, J. and Casas, A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annual Review of Political Science* 20:529–544.

Wilkinson, P. (2007). *International Relations: A Very Short Introduction.* Oxford, United Kingdom: Oxford University Press.

Williams, C. C., Allison, J. G., Vidal, G. A., Burow, M. E., Beckman, B. S., Marrero, L. and Jones, F. E. (2004). The ERBB4/HER4 receptor tyrosine kinase regulates gene expression by functioning as a STAT5A nuclear chaperone. *Journal of Cell Biology* 167:469–478.

Willnat, L. and Luo, Y. (2011). Watching the dragon: global television news about China. *Chinese Journal of Communication* 4:255–273.

Xiang, D. (2013). China's image on international English language social media. *The Journal of International Communication* 19:252–271.

Xie, T. and Page, B. I. (2013). What affects China's national image? A cross-national study of public opinion. *Journal of Contemporary China* 22:850–867.

Yang, W. (2005). *Design of a knowledge acquisition tool using a constructivist approach for creating tailorable patient education materials.* Master's thesis. Waterloo, Canada: University of Waterloo.

Yao, J.-g., Wan, X. and Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems* 53:297–336.

Young, L. and Soroka, S. (2012). Affective news: the automated coding of sentiment in political texts. *Political Communication* 29:205–231.

Zalon, M. L., Sandhaus, S., Kovaleski, M. and Roe-Prior, P. (2017). Hospitalized older adults with established delirium: recognition, documentation, and reporting. *Journal of Gerontological Nursing* 43:32–40.

Zerva, C. and Ananiadou, S. (2018). Paths for uncertainty: exploring the intricacies of uncertainty identification for news. In Blanco, E. and Morante, R. (eds.), *Proceedings of the Workshop on Computational Semantics beyond Events and Roles* (pp. 6–20). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zhang, J., Spirling, A. and Danescu-Niculescu-Mizil, C. (2017). Asking too much? The rhetorical role of questions in political discourse. In Palmer, M., Hwa, R. and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1558–1572). Stroudsburg, PA, USA: Association for Computational Linguistics.

Part II

Original publications

Publication I

Heimonen, J., Pyysalo, S., Ginter, F. and Salakoski, T. (2008). Complex-to-pairwise mapping of biological relationships using a semantic network representation. In Salakoski, T., Rebholz-Schuhmann, D. and Pyysalo, S. (eds.), *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (pp. 45–52). Turku, Finland: Turku Centre for Computer Science

URL: `https://tucs.fi/publications/view/?pub_id=inpHePyGiSa08a`
A reprint is only available in the printed version.

Publication II

Heimonen, J., Björne, J. and Salakoski, T. (2010). Reconstruction of semantic relationships from their projections in biomolecular domain. In Cohen, K. B., Demner-Fushman, D., Ananiadou, S., Pestian, J., Tsujii, J. and Webber, B. (eds.), *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* (pp. 108–116). Stroudsburg, PA, USA: Association for Computational Linguistics

URL: `https://www.aclweb.org/anthology/W10-1914/`
A reprint is only available in the printed version.

Publication III

Heimonen, J., Salakoski, T. and Pahikkala, T. (2014). Properties of object-level cross-validation schemes for symmetric pair-input data. In Fränti, P., Brown, G., Loog, M., Escolano, F. and Pelillo, M. (eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 384–393). Berlin Heidelberg, Germany: Springer

URL: https://doi.org/10.1007/978-3-662-44415-3_39

A reprint is only available in the printed version.

Publication IV

Heimonen, J., Danielsson-Ojala, R., Salakoski, T., Lundgrén-Laine, H. and Salanterä, S. (2018). Ontology development for patient education documents using a professional- and patient-oriented Delphi method. *CIN: Computers, Informatics, Nursing* 36:448–457

URL: `https://doi.org/10.1097/CIN.0000000000000436`
A reprint is only available in the printed version.

Publication V

Moen, H., Peltonen, L.-M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T. and Salanterä, S. (2016). Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine* 67:25–37

URL: `https://doi.org/10.1016/j.artmed.2016.01.003`

A reprint is only available in the printed version.

Publication VI

Aukia, J., Heimonen, J., Pahikkala, T. and Salakoski, T. (2017). Automated quantification of Reuters news using a receiver operating characteristic curve analysis: the Western media image of China. *Global Media and China* 2:251–268

URL: `https://doi.org/10.1177/2059436418754890`
A reprint is only available in the printed version.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

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

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönnqvist**, Knowledge-Lean Text Mining
228. **Mohammad-Hashem Hahgbayan**, Energy-Efficient and Reliable Computing in Dark Silicon Era
229. **Charmi Panchal**, Qualitative Methods for Modeling Biochemical Systems and Datasets: The Logicome and the Reaction Systems Approaches
230. **Erkki Kaila**, Utilizing Educational Technology in Computer Science and Programming Courses: Theory and Practice
231. **Fredrik Robertsén**, The Lattice Boltzmann Method, a Petaflop and Beyond
232. **Jonne Pohjankukka**, Machine Learning Approaches for Natural Resource Data
233. **Paavo Nevalainen**, Geometric Data Understanding: Deriving Case-Specific Features
234. **Michal Szabados**, An Algebraic Approach to Nivat's Conjecture
235. **Tuan Nguyen Gia**, Design for Energy-Efficient and Reliable Fog-Assisted Healthcare IoT Systems
236. **Anil Kanduri**, Adaptive Knobs for Resource Efficient Computing
237. **Veronika Suni**, Computational Methods and Tools for Protein Phosphorylation Analysis
238. **Behailu Negash**, Interoperating Networked Embedded Systems to Compose the Web of Things
239. **Kalle Rindell**, Development of Secure Software: Rationale, Standards and Practices
240. **Jurka Rahikkala**, On Top Management Support for Software Cost Estimation
241. **Markus A. Whiteland**, On the k-Abelian Equivalence Relation of Finite Words
242. **Mojgan Kamali**, Formal Analysis of Network Routing Protocols
243. **Jesús Carabaño Bravo**, A Compiler Approach to Map Algebra for Raster Spatial Modeling
244. **Amin Majd**, Distributed and Lightweight Meta-heuristic Optimization Method for Complex Problems
245. **Ali Farooq**, In Quest of Information Security in Higher Education Institutions: Security Awareness, Concerns, and Behaviour of Students
246. **Juho Heimonen**, Knowledge Representation and Text Mining in Biomedical, Healthcare, and Political Domains

Turku Centre *for* Computer Science

http://www.tucs.fi

tucs@abo.fi

University of Turku
Faculty of Science and Engineering
- Department of Future Technologies
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3881-9
ISSN 1239-1883