Doctoral theses at NTNU, 2016:134

Hans Moen

Distributional Semantic Models for Clinical Text Applied to Health Record Summarization

Doctoral thesis

Hans Moen

ISBN 978-82-326-1606-0 (printed ver.)
ISBN 978-82-326-1607-7 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of Science and Technology

Turun yliopisto
University of Turku

NTNU

Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science
University of Turku
Faculty of Mathematics and Natural Sciences
Department of Information Technology

Doctoral theses at NTNU, 2016:134

NTNU

NTNU
Norwegian University of Science and Technology

Hans Moen

Distributional Semantic Models for Clinical Text Applied to Health Record Summarization

Thesis for the Degree of Philosophiae Doctor

Trondheim, May 2016

Norwegian University of Science and Technology
Faculty of Information Technology,
Mathematics and Electrical Engineering
Department of Computer and Information Science

University of Turku
Faculty of Mathematics and Natural Sciences
Department of Information Technology

NTNU
Norwegian University of
Science and Technology

Supervisors

Norwegian University of Science and Technology, Norway

Associate Professor Øystein Nytrø, PhD
Department of Computer and Information Science

Professor Björn Gambäck, PhD
Department of Computer and Information Science

Associate Professor Pinar Öztürk, PhD
Department of Computer and Information Science

University of Turku, Finland

Professor Sanna Salanterä, PhD, RN
Department of Nursing Science

Professor Tapio Salakoski, PhD
Department of Information Technology

Others
Doctor Laura Slaughter, Senior Researcher, PhD
The Intervention Centre, Oslo University Hospital, Norway

Jo nt de ree et een:

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering Department of Computer and Information Science

Un ers t of T rk
F lt o M t em t s n N t r l S en es
Dep rtment o In orm t on Te nolo

T es s or t e De ree o P losop eDo tor

ISBN 978-82-326-1606-0 (printed ver.)
ISBN 978-82-326-1607-7 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2016:134

Printed by NTNU Grafisk senter

Abstract

As information systems in the health sector are becoming increasingly computerized, large amounts of care-related information are being stored electronically. In hospitals clinicians continuously document treatment and care given to patients in electronic health record (EHR) systems. Much of the information being documented is in the form of clinical notes, or narratives, containing primarily unstructured free-text information. For each care episode, clinical notes are written on a regular basis, ending with a discharge summary that basically summarizes the care episode. Although EHR systems are helpful for storing and managing such information, there is an unrealized potential in utilizing this information for smarter care assistance, as well as for secondary purposes such as research and education. Advances in clinical language processing are enabling computers to assist clinicians in their interaction with the free-text information documented in EHR systems. This includes assisting in tasks like query-based search, terminology development, knowledge extraction, translation, and summarization.

This thesis explores various computerized approaches and methods aimed at enabling automated semantic textual similarity assessment and information extraction based on the free-text information in EHR systems. The focus is placed on the task of (semi-)automated summarization of the clinical notes written during individual care episodes. The overall theme of the presented work is to utilize resource-light approaches and methods, circumventing the need to manually develop knowledge resources or training data. Thus, to enable computational semantic textual similarity assessment, word distribution statistics are derived from large training corpora of clinical free text and stored as vector-based representations referred to as distributional semantic models. Also resource-light methods are explored in the task of performing automatic summarization of clinical free-text information, relying on semantic textual similarity assessment. Novel and experimental methods are presented and evaluated that focus on: a) distributional semantic models trained in an unsupervised manner from statistical information derived from large unannotated clinical free-text corpora; b) representing and computing semantic similarities between linguistic items of different granularity, primarily words, sentences and clinical notes; and c) summarizing clinical free-text information from individual care episodes.

Results are evaluated against gold standards that reflect human judgements. The results indicate that the use of distributional semantics is promising as a resource-light approach to automated capturing of semantic textual similarity relations from unannotated clinical text corpora. Here it is important that the semantics correlate with the clinical terminology, and with various semantic similarity assessment

tasks. Improvements over classical approaches are achieved when the underlying vector-based representations allow for a broader range of semantic features to be captured and represented. These are either distributed over multiple semantic models trained with different features and training corpora, or use models that store multiple sense-vectors per word. Further, the use of structured meta-level information accompanying care episodes is explored as training features for distributional semantic models, with the aim of capturing semantic relations suitable for care episode-level information retrieval. Results indicate that such models performs well in clinical information retrieval. It is shown that a method called Random Indexing can be modified to construct distributional semantic models that capture multiple sense-vectors for each word in the training corpus. This is done in a way that retains the original training properties of the Random Indexing method, by being incremental, scalable and distributional. Distributional semantic models trained with a framework called Word2vec, which relies on the use of neural networks, outperform those trained using the classic Random Indexing method in several semantic similarity assessment tasks, when training is done using comparable parameters and the same training corpora. Finally, several statistical features in clinical text are explored in terms of their ability to indicate sentence significance in a text summary generated from the clinical notes. This includes the use of distributional semantics to enable case-based similarity assessment, where cases are other care episodes and their "solutions", i.e., discharge summaries. A type of manual evaluation is performed, where human experts rates the different aspects of the summaries using a evaluation scheme/tool. In addition, the original clinician-written discharge summaries are explored as gold standard for the purpose of automated evaluation. Evaluation shows a high correlation between manual and automated evaluation, suggesting that such a gold standard can function as a proxy for human evaluations.

Preface

The research was part of a four-year PhD program at the Department of Computer and Information Science, Faculty of Information Technology, Mathematics and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Norway. During this period, 25 per cent of the time was devoted to teaching duties at NTNU. The research was also part of a cotutelle collaboration with the Department of Information Technology, University of Turku (UTU), Finland, where I spent a total of one year. The PhD period lasted approximately five years in total.

The research at NTNU was part of, and funded by, a national project called Evidence-based Care Processes: Integrating Knowledge in Clinical Information Systems (EviCare). This was financed by the Research Council of Norway (NFR), project no. 193022.

The stay and research at UTU was partly funded by the Academy of Finland, project no. 140323. At UTU I participated in a consortium named Information and Language Technology for Health Information and Communication (IKITIK[1]).

During the first three years of the PhD period, I participated in a Nordic and Baltic collaboration network named Health Text Analysis Network in the Nordic and Baltic Countries (HEXAnord[2]), funded by Nordforsk (Nordic Council).

[1] `http://www.utu.fi/en/units/med/units/hoitotiede/research/projects/ikitik` (accessed 1st March 2016)

[2] `http://dsv.su.se/en/research/research-areas/health/hexanord-1.113078` (accessed 1st March 2016)

Acknowledgements

First I want to thank Øystein Nytrø for providing me with the opportunity to pursue the PhD degree and for his guidance in the process. Then I want to thank Björn Gambäck for his thorough and inspiring supervision and collaboration. I am grateful to Pinar Öztürk for her advice, for introducing me to the topic of distributional semantics and for providing me with the foundation and inspiration to start the PhD studies through her supervision of my Master's thesis. Thanks also go to Laura Slaughter for her help and advice.

A special and generous thanks to Erwin Marsi at NTNU for his invaluable support, supervision, collaboration and inspiration through the entire process, despite not being my supervisor on paper.

I want to thank Sanna Salanterä and Tapio Salakoski for inviting me to come to UTU, for the warm welcome provided for me there, and for the following supervision and collaboration, which has been a very important part of my PhD studies. I also would like to thank the others affiliated with UTU that I have collaborated with in terms of co-authorship. These are primarily Filip Ginter, Laura-Maria Peltonen (formerly Murtola), Antti Airola, Tapio Pahikkala, Juho Heimonen, Sampo Pyysalo, Heljä Lundgren-Laine, Riitta Danielsson-Ojala and Virpi Terävä.

I would also like to acknowledge the sources of funding that made this work possible: The Research Council of Norway (NFR) which funded the EviCare project; The Department of Computer and Information Science at NTNU; The Academy of Finland, University of Turku; Nordforsk (Nordic Council) which funded the HEXAnord project.

I want to thank those involved in the EviCare project. I am grateful to Thomas Brox Røst for early on in the process providing me with useful and street smart advice about being a PhD student.

The HEXAnord project was a excellent forum for meeting other researchers in the Nordic region with similar interest areas. I was fortunate to establish a close connection to many of the involved partners. This again resulted in international research collaboration with several partners. I want to thank everyone involved in the HEXAnord project, in particular Aron Henriksson, Maria Skeppstedt and Martin Duneld (formerly Hassel) (Stockholm University, Sweden), Vidas Daudaravičius (Vytautas Magnus University, Lithuania) and Ann-Marie Eklund (University of Gothenburg, Sweden) for the productive collaboration and co-authorship. I also like to thank Marie Lindbergh (Linköping University, Sweden) for the collaboration in the HEXAnord PhD courses. Thanks also go to Hercules Dalianis (Stockholm University, Sweden) for his work as the project manager.

I am grateful for the assistance from Jari Björne (UTU) for his advice and input during the writing of the PhD thesis and Stewart Clark (NTNU) who assisted with the final editing.

I want to thank my office mates at NTNU, Simone Mora, Elena Parmiggiani and Francesco Gianni for including me in the Italian society there. Another generous thanks go to the "lunch people" at NTNU, primarily being Gleb Sizov, Erik Smistad, Kai Olav Ellefsen, Lars Bungum, Axel Tidemann, Lester Solbakken and Boye Annfelt Høverstad for all the interesting and entertaining discussions. At the Department of Nursing Science at UTU I shared office with Eriikka Siirala and Hannakaisa Niela-Vilén who I would like to thank for contributing to a nice work atmosphere.

I am grateful to my friends who have played an important role through these years. Special thanks go to Tove Sveen for her helpful suggestions and to Kevin Andre Vatn for contributing to solving certain hardware- and Java-related issues.

Finally, and most importantly, I would like to thank my parents, Jorunn Iversen and Vidar Moen, and my siblings, Martin Moen, Olve Moen and Kjersti Moen, for their constant support, love and encouragement.

Contents

Part I

Research Overview

Chapter 1

Introduction

The work conducted in this thesis approaches the task of automated summarization of clinical free text in care episodes. Focus is placed on methods that mainly exploit distributional statistics in clinical notes and care episodes, thus avoiding manual labor in constructing semantic knowledge resources to support this task. A set of different distributional semantic methods, i.e the models they construct, are first evaluated in the following separate tasks: *synonym extraction* (word similarity assessment), *sentence similarity classification* (sentence similarity assessment), and *care episode retrieval* (care episode similarity assessment). Each of these represents tasks related to supporting clinical work, while also directly or indirectly representing sub-tasks in a intended text summarization system. Finally these models are used in a set of methods for performing *automatic summarization of care episodes*. The work touches upon a number of fields related to *natural language processing*, primarily *computational semantics*, *information retrieval* and *automatic text summarization*.

1.1 Motivation

The development, adoption and implementation of health information technology, such as *electronic health record* (EHR) systems, are strategic focuses of health policies globally European Commission (2012), Blumenthal and Tavenner (2010), Jha (2010), Bartlett et al. (2008). The amount of electronically documented health information is increasing as health records are becoming computerized. In addition, the ongoing advances in diagnostic and health sciences contribute to an increase in the amount of information accumulated for each patient. The large amounts of computerized health information complicate its management and increase the risk of information overload for clinicians (Hall and Walton 2004, Farri

et al. 2012). This causes other problems in the clinical work, such as errors, frustration, inefficiency, and communication failures (Lissauer et al. 1991, Suominen and Salakoski 2010). At the same time, this creates opportunities for technological solutions to support clinical care and research.

In hospitals, much of the information that clinicians document are notes in the form of free text that they write in relation to patient care. During a patient's hospital stay, i.e., a *care episode*, clinicians with various specializations write *clinical notes* on a regular basis to document the ongoing care process (status, reasoning, plans, findings, operations, etc.). In the end, normally when the patient is leaving the hospital, a *discharge summary* is written that summarizes the hospital stay. The free text in clinical notes may contain valuable information that is not found or documented elsewhere, such as in the structured, numerical and image data stored in EHRs.

Natural language processing (NLP) (Hirschberg and Manning 2015) tools and resources have the potential to assist clinicians in their interaction with this free-text information. This includes assisting in tasks like automatic event detection in health records (Mendonça et al. 2005), automatic concept indexing (Berman 2004), medication support (Xu et al. 2010), decision support (Demner-Fushman et al. 2009, Velupillai and Kvist 2012), query-based search (Grabar et al. 2009)) and automated summarization (Pivovarov and Elhadad 2015).

These tasks require a certain amount of understanding of the "meaning" of the linguistic items, such as words, sentences and documents. Here, methods in *computational semantics* can be used that focus on how to automate the process of constructing and reasoning with meaning representations of linguistic items. An active area in computational semantics focuses on methods for doing automated *semantic similarity* assessment, which utilizes a similarity metric to calculate a numeric value reflecting the likeness of the meaning, or semantic content, between pairs of linguistic items.

Clinical language has a highly domain-specific terminology, thus specialized NLP tools and resources are commonly used to enable computerized analysis, interpretation and management of written clinical text (Kvist et al. 2011, Meystre et al. 2008, Pradhan et al. 2014). This includes tasks involving computerized semantic similarity assessment. As an example, we have the following two sentences, both referring to the same event and patient, written by two different clinicians:

- *"The patient has broken his right foot during a football match."*
- *"Pt fractured his right ankle when playing soccer."*

This example illustrates how two sentences that barely contain any of the same words can describe one and the same event. A straightforward string matching approach is not adequate to determine that they have a similar meaning, thus a more advanced approach is needed.

There are several lexical resources that enable various degrees of computational semantic textual similarity assessment to be made between words and concepts found in the clinical terminology. Examples of such resources are: the Systematized Nomenclature of Medicine, Clinical Terms (SNOMED-CT) ontology (NLM b); the Medical Subject Headings (MeSH) thesaurus (NLM a); the International Classification of Diseases[1] (ICD) medical classification lists (World Health Organization 1983); the Unified Medical Language System (UMLS) compendium (NLM c) where all these resources are part of or originated from; the generic WordNet ontology (Miller 1995).

Unfortunately, the above lexical resources exist primarily for the English language and/or have limited generalizability in terms of language and coverage, which limits their area of use. Developing new resources, or adapting existing resources to other languages, is costly and time consuming as it requires labor by experts with both linguistic and domain knowledge — medical and clinical knowledge. Such efforts are often done through extensive national or international collaboration projects, one example being the translation of MeSH into Norwegian (Aasen 2012). Further, even though thesauri and ontologies (and such a manual modeling approach in general) are well suited for modeling (semantic) relations on a conceptual level, modeling all possible semantic relations between, e.g., words and concepts used in clinical text would be very difficult and costly to achieve. To enable fine-grained computerized semantic similarity assessment between all possible linguistic items found in clinical text, one would need to develop or adapt semantic resources to the point where they (in sum) capture the totality of the localized terminology used by clinicians in the language(s), region(s), hospital(s) and ward(s) of interest. On top of this come the potential problems related to legal restrictions in terms of distributing clinical information due to its potentially sensitive content. This limits the number of researchers and developers accessing relevant data in the first place. In addition, this is a limiting factor with respect to the amount and coverage of openly available clinical language resources relevant to enable semantic similarity assessment.

An alternative approach focuses on enabling automated, data-driven, learning of semantic similarity in the vocabulary in a text corpus. Such methods are commonly referred to as *distributional semantic methods* (see Turney and Pantel (2010) for

[1] International Statistical Classification of Diseases and Related Health Problems

an overview). Central here is the process of capturing semantic relations based on statistics about word usage in the corpus, and storing this as a computerized vector-based representation, typically referred to as a model — a *distributional semantic (similarity) model*. In applying such methods one can potentially circumvent the need to manually develop lexical resources, in particular those focusing on semantic similarity, or reduce the need for manual labour through hybrid approaches. Pedersen et al. (2007) showed that distributional semantic methods — that exploit statistical distribution patterns in unannotated, unstructured, free text in a *training corpus* of clinical text — are suited for the modeling of semantic similarities between medical concepts on the same level as using SNOMED-CT, WordNet and other available resources. These methods rely on the *distributional hypothesis* (Harris 1954) (see Section 2.1), and their underlying representations are vector-based — vector space models (VSMs) (see Section 2.1.3). They produce distributional semantic models, which are also referred to as *distributional semantic spaces*, in the form of a vector-based representation that enables the computer to calculate similarity between linguistic items (e.g., words, sentences, documents) as a distance measure. Training of such models is commonly done using an unannotated, unstructured, free-text corpora, thus this type of methods can be said to be language independent and "resource light". As the training is data-driven, the resulting models tend to reflect the semantic relations in the language and terminology used in the utilized training corpus. However, there are numerous ways of constructing distributional semantic models with respect to what features to use for training, how to weight the features, how to represent the semantic information, how to calculate similarities between the constituent vectors (i.e., what similarity metric to use), and so on. This necessitates exploration of various ways of capturing and calculating the desired semantics from a training corpus that best match the similarity assessment task at hand (see, e.g., Kolb (2009), Baroni and Lenci (2010), Lenci and Benotto (2012)).

A possible application is related to the discharge summaries that clinicians write when summarizing patients' hospitalization periods (i.e., care episodes). Due to factors such as limited time and information overload, discharge summaries are often produced late, and the information they contain tends to be insufficient (Kripalani et al. 2007). Thus, clinicians would potentially benefit from having a system that supports information summarization through (semi-)automatic text summarization, not only during the discharge process, but also at any point during an ongoing care episode, and for summarizing information from earlier care episodes (Pivovarov and Elhadad 2015). Ultimately such a system could help in saving time and improving the quality of documentation in hospitals. Figure 1.1 illustrates a care episode consisting of several clinical notes, and ends with a discharge summary.

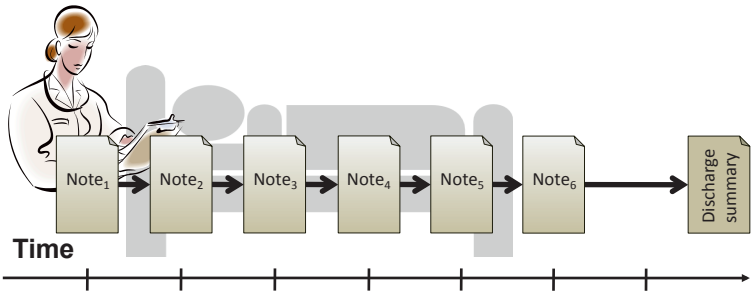

Figure 1.1: A care episode, consisting of a set of clinical notes and ending with a discharge summary.

Automatic text summarization is the computerized process of taking some text from one or more documents and constructing a shortened version that retains the most important information (Luhn 1958). The task of summarizing multiple clinical notes from one care episode is a matter of summarizing multiple documents — i.e., *multi-document summarization* — involving the following goals: include the most important or relevant information; avoid redundant information; produce coherent text. There are many ways to approach this task. (Jones 1999) presents *factors* that one has to be taken into account in order to make a summarization system achieve its task. These mainly concerns *input*, *purpose* and *output*. Others have later discussed and elaborated upon these factors Hahn and Mani (2000), Afantenos et al. (2005). Through a study conducted early on in the PhD process, we identified the following properties and requirements for a text summarization system intended for clinical free-text notes (see Section 1.2): It concerns *multiple documents*; few tailored lexical and knowledge *resources* exist; the content selection is to be done in an *extraction-based* fashion; the produced summaries should contain *indicative* information; the system should be able to produce both *generic* and *user-oriented* summaries. The *output* should be a single piece of text, with similar structure as the notes written by clinicians, which would arguably make *evaluation* more convenient compared to other alternatives, such as graph- or time-line-based visualization. See Section 2.3 for more details.

Selecting what information to include when summarizing the textual content in a care episode is a complex and challenging task. A recent review by Mishra et al. (2014) found that most text summarization techniques used in the biomedical domain can be classified as "knowledge rich" as they depend on (and the quality of) manually developed lexical resources, such as ontologies and annotated training

corpora and gold standards[2] available for training and testing. The same seems to apply to techniques and methods in existing summarization systems designed for EHRs (Pivovarov and Elhadad 2015). Typically such knowledge resources are used to first explicitly classify the information in the text that is to be summarized, such as words and concepts, and then used to assess similarities and ultimately significance. This however implies that the systems have restricted generalizability in terms of languages and (sub-)domains.

Textual similarity assessment, particularly on a sentence level, is an important aspect of automatic text summarization (Ferreira et al. 2016). Pivovarov and Elhadad (2015) observed that, in clinical summarization, there has been relatively little work on similarity identification between textual concepts in (sequences of) clinical notes, including the exploration of such information for the purpose of automated summarization. Further, this is identified as an important direction for future EHR summarization methodology. In the approach presented in this thesis (Paper E), a set of techniques and methods are explored and evaluated that uses various types of statistically derived information and features found within the care episode that are to be summarized, and/or in large collections of care episodes. Although this makes the methods/techniques arguably "knowledge poor", they are easily adaptable to different languages and sub-domains within the health sector. One example is to explore various textual features found internally in a care episode, such as word usage and repeated information; another example is to look at other care episodes with similar content, selected using *information retrieval* (Manning et al. (2008), Chapter 6) (Paper C), and then look at the statistical probability for some information to be found in a discharge summary given that it occurs in one or more of its accompanying clinical notes. These examples depend upon the ability to measure semantic similarity between linguistic items, such as words sentences and documents, which motivates the use of *distributional semantics* (Paper A, B and C).

1.2 Research Objectives

The present research was driven by a set of goals (RG1–RG4), each leading to the next. The initial goal (RG1) was provided by the EviCare project:

RG1: *Explore approaches for conducting summarization of the free text in care episodes, emphasizing approaches and underlying methods that are resource light in terms of adaptation to the domain and different languages.*

[2] A *gold standard*, or *reference standard*, is here defined as being the optimal/ideal solution for the task at hand.

This led to:

RG2: *Explore various (vector-based) distributional semantic methods with respect to their ability to capture semantic similarity between linguistic items in clinical (free) text.*

This again led to:

RG3: *Explore ways to enable distributional semantic methods/models to capture domain- and task-specific semantic information from clinical free text, focusing on the following two tasks:*

- *Sentence similarity classification.*
- *Care episode similarity assessment.*

When pursuing the above goals, conducting a proper evaluation became yet another goal:

RG4: *Find how to automatically and reliably evaluate the various text summarization approaches and the underlying semantic methods in the sub-tasks they are intended for.*

With these goals in mind, RG1 in particular, I had the following research questions that I intended to answer through a set of experiments:

RQ1: *How can the distributional hypothesis be utilized in constructing semantic similarity models suited for clinical text?*

RQ2: *What sentence-level features of clinical text in care episodes are indicative of relevancy for inclusion in a clinical free-text summary?*

RQ3: *How can the evaluation of distributional semantic models and text summaries generated from clinical text be done in a way that is fast, reliable and inexpensive?*

In the work on addressing these research questions, four sets of experiments were conducted. The first set focuses on synonym extraction, the second concerns sentence similarity classification, then care episode retrieval, and finally automatic summarization of care episodes. The clinical text used is mainly from a Swedish

and a Finnish hospital, as detailed in Section 1.4.4. The utilized methods are considered language independent (when not taking into consideration the text lemmatization), but are arguably somewhat biased towards the clinical documentation procedures and structure that is common in this region (Allvin et al. 2010). These sets of experiments are presented in five separate papers, as shown in Table 1.1. They build on each other and Figure1.2 visualizes these relations, starting from word-level similarity assessment. The relations between *Research Goals*, *Research Questions* and *Papers* is shown in Table 1.2.

Experiments	Papers
Synonym extraction	A: *Synonym extraction and abbreviation expansion with ensembles of semantic spaces*
Sentence similarity classification	B: *Towards dynamic word sense discrimination with Random Indexing*
Care episode retrieval	C: *Care episode retrieval: distributional semantic models for information retrieval in the clinical domain*
Automatic summarization of care episodes (1)	D: *On evaluation of automatically generated clinical discharge summaries*
Automatic summarization of care episodes (2)	E: *Comparison of automatic summarization methods for clinical free-text notes*

Table 1.1: Experiments and accompanying papers.

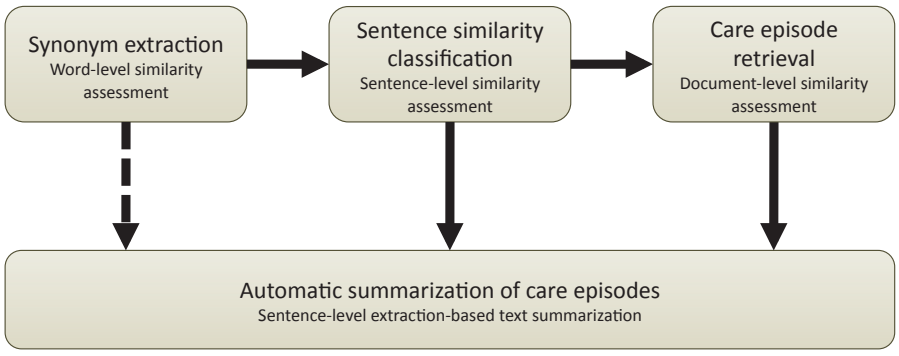

Figure 1.2: An overview of the conducted research.

1.3 Research Methodology

This thesis work touches upon a number of fields related to NLP. Primarily these are computational semantics, information retrieval and automatic text summarization. As most of the tasks and experiments directly or indirectly focuses on sup-

Research Goals	Research Questions	Papers
1	1, 2	A, B, C, D, E
2	1	A
3	1, 3	B, C
4	3	A, C, D, E

Table 1.2: The relation between *Research Goals*, *Research Questions* and *Papers*.

porting health care, this work is also related to the field of health informatics. The work also involved various degrees of collaboration with clinical professionals, which provided invaluable insight and understanding of their work practice and needs. Innovation was a keyword for the EviCare project and the IKITIK consortium, which is also reflected in this work.

The overall research approach can be viewed as *design science* (Hevner et al. 2004). In the design science paradigm the aim is to design and apply new and/or innovative artifacts aimed at human and organizational use. Knowledge and understanding about the underlying domain and possible solutions are gained through the design, application and evaluation process of the artifact(s), often performed in iterations. As emphasized by Cohen and Howe (1988; 1989), artificial intelligence research should be driven by evaluation. When developing a system or program in this field, evaluation should not only cover performance measures, but also reveal the behavior of the system, limitations, generalizability and prospects for future development.

Starting from RG1, the general direction of the research was set relatively early on in the process. The research questions were then defined in the process of deciding on a general level what techniques and methods that I wanted to explore when approaching RG1. From there the various research goals following RG1 emerged. The various techniques and methods utilized in the different experiments reflects the underlying hypotheses.

Primarily an iterative process was used when conducting the experiments, where each iteration typically included design (software design and implementation), application and evaluation. Mainly a *quantitative* approach (see, e.g., VanderStoep and Johnson (2008), page 7) was used for evaluation. Performance scores were calculated based on gold standards and further compared to scores achieved by various related approaches (baselines and state-of-the-art). In that sense the process was guided by the gold standards used in the various experiments. Through analysing the evaluation scores and identifying problems that arose during the implementation and application, increased understanding was gained regarding the

utilized methods in terms of their potential applications, strengths and weaknesses. When developing the manual evaluation scheme related to the automatic text summarization work, the use of open-ended questions was also explored. The latter provided some *qualitative* feedback (see, e.g., VanderStoep and Johnson (2008), page 7) from clinical experts about the direction of that work.

The presented results and methods could potentially contribute to approaches and software methods for others to use and expand upon when pursuing similar goals. Results of this work have been published in conference/workshop proceedings and journals. The experiments and utilized resources are explained in ways that should enable others to replicate the experiments. However, the clinical corpora used are not openly available due to the sensitive nature of clinical text.

1.4 Research Papers and Contributions

1.4.1 List of Papers Included in the Thesis

Paper A: Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Daudaravičius, Vidas, and Duneld, Martin. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):25, 2014.

Paper B: Moen, Hans; Marsi, Erwin, and Gambäck, Björn. Towards dynamic word sense discrimination with Random Indexing. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 83–90, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Paper C: Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

Paper D: Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014. CEUR Workshop Proceedings.

Paper E: Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna. Comparison of automatic summarisation methods for clinical free text notes. *Artificial In-*

telligence in Medicine, 67:25–37, 2016.

1.4.2 List of Related Papers Not Directly Included in the Thesis

Öztürk, Pinar; Prasath, R. Rajendra, and Moen, Hans. Distributed representations to detect higher order term correlations in textual content. In *Rough Sets and Current Trends in Computing - 7th International Conference, RSCTC 2010*, volume 6086 of *Lecture Notes in Computer Science*, pages 740–750, Warsaw, Poland. Springer, 2010.

Moen, Hans and Marsi, Erwin. Towards retrieving and ranking clinical recommendations with Cross-lingual Random Indexing. In *Proceedings of CLEFeHealth 2012, CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, volume 1178 of *CEUR Workshop Proceedings*, numpages 4, Rome, Italy, 2012.

Henriksson, Aron; Moen, Hans; Skeppstedt, Maria; Eklund, Ann-Marie; Daudaravičius, Vidas, and Hassel, Martin. Synonym extraction of medical terms from clinical text using combinations of word space models. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012)*, pages 10–17, Zurich, Switzerland, 2012. Zurich Open Repository and Archive.

Moen, Hans and Marsi, Erwin. Towards cross-lingual information retrieval using Random Indexing. In *NIK: Norsk Informatikkonferanse, volume 2012*, pages 259–262, Bodø, Norway, 2012. Akademika forlag.

Marsi, Erwin; Moen, Hans; Bungum, Lars; Sizov, Gleb; Gambäck, Björn, and Lynum, André. NTNU-CORE: Combining strong features for semantic similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 66–73, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Moen, Hans and Marsi, Erwin. Cross-lingual Random Indexing for information retrieval. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 164–175. 2013.

Murtola, Laura-Maria; Moen, Hans; Kauhanen, Lotta; Lundgrén-Laine, Heljä; Salakoski, Tapio, and Salanterä, Sanna. Using text mining to explore concepts associated with acute confusion in cardiac patients documentation. In *Proceedings of CLEFeHealth 2013: Student Mentoring Track*, numpages 2, Valencia, Spain, 2013. CLEF online working notes.

Pyysalo, Sampo; Ginter, Filip; Moen, Hans; Salakoski, Tapio, and Ananiadou, Sophia. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LMB 2013)*, pages – 5, Tokyo, Japan, 2013. Database Center for Life Science.

Moen, Hans; Marsi, Erwin; Ginter, Filip; Murtola, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi 2014) @ EACL*, pages 116–124, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

1.4.3 Contributions

The main contributions of the work presented in the thesis are:

- Paper A: Evaluation of several vector-based distributional semantic models for automated synonym extraction from clinical text, including exploration of combined model pairs trained on clinical text or medical journal articles.
- Paper B: Introduction of a novel method for constructing multi-sense semantic models, evaluated in a task concerning sentence similarity assessment.
- Paper C: Evaluation of a set of information retrieval methods that utilize the distributional hypothesis. The resulting models are evaluated in the task of care episode retrieval. These experiments include novel methods utilizing the ICD-10 codes attached to care episodes to better induce domain-specificity in the resulting models.
- Papers A & C: Proposals for how to evaluate semantic models used for clinical synonym extraction and care episode similarity.
- Paper E: Exploration of a set of resource-light automatic text summarization methods tailored for sequences of clinical (free-text) notes in care episodes.
- Papers D & E: Proposals for how to evaluate clinical text summaries; in an automatic and manual way.

1.4.4 Clinical Corpora

It is typically very difficult for researchers to enquire access to collections of personal health documents of significant size. An asset in the present work is that relatively large corpora of clinical text are used in the experiments.

The corpus used in Paper A is a subset of the *Stockholm EPR Corpus* (Dalianis et al. 2009), extracted from a Swedish hospital and contains clinical notes written primarily in Swedish by physicians, nurses and other health care professionals over a period of six months. It consists of 268 727 notes and approximately 42.5 million words. The use of this corpus for research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5. In Papers C, D and E the corpus used is extracted from a Finnish hospital, over a period of four years, consisting of clinical notes written in primarily Finnish by physicians for patients with any type of heart-related problems. It consists of 398 040 notes and approximately 64 million words. Ethical approval for the research was obtained from the Ethics Committee of the Hospital District (17.2.2009 §67), and permission to conduct the research was obtained from the Medical Director of the Hospital District, permission number 2/2009. These corpora are stored in compliance to local regulations concerning sensitive data management.

1.5 Thesis Structure

The remainder of the thesis is structured as follows.

- **Chapter 2** introduces the main concepts and methods that are needed in order to understand the work in the included papers.
- **Chapter 3** contains the main results in the form of paper summaries and retrospective discussions.
- **Chapter 4** discusses the main contributions in relation to the research questions and discusses some directions for future work.
- **Part II** contains the papers of the thesis.

Chapter 2

Background

This chapter provides an overview of the related background for the work presented in the thesis.

2.1 Computational Semantics and Distributional Semantics

Human language is very complex as it reflects high-level cognitive processes of the human brain. To fully understand the true or intended "meaning" and/or content of a word, sentence or document, one needs understanding and knowledge about: The language and underlying grammar and syntax; The meaning of each word and what information they are meant to convey, alone and in context with other words and word phrases, posterior and anterior ones; How each word and word phrase relates to concepts or objects in the real world; The domain, topic and ongoing event(s) or case(s). Even the concept "meaning" itself is rather defuse, as elaborated by Sahlgren (2006).

Tasks or problems requiring *artificial intelligence* (AI) for solving on the same level as human intelligence are commonly referred to as being "AI-complete" (Yampolskiy 2013)[1]. On the one hand, there is a large gap between the cognitive processes underlying language understanding native to human intelligence and that which is achieved by today's computers. On the other hand, many tasks involving processing of natural language do not necessary require a deep understanding of the information it is meant to convey. Today we see that rather shallow approaches can be of great assistance in multiple *natural language processing* (NLP) tasks — approaches that exploit the computational power of computers and the exis-

[1] An AI-complete problem means that its optimal solution implies solving the problem of developing AI in computers that are as intelligent as humans.

tence of large amounts of accumulated digital (textual) information (Hirschberg and Manning 2015). This is reflected in state-of-the-art machine translation systems, search engines and question answering systems, e.g., IBM's Watson system (Ferrucci et al. 2010).

NLP is a field that concerns the interaction between computers and human natural languages (Hirschberg and Manning 2015). An example of a computer system that includes NLP is one that takes human language as input, written or spoken, then tries to interpret and "understand" it in a way by, e.g., converting it into a representation that the computer can further process. This can be to convert free-text queries into a form or representation in an Internet search engine that is used to find web pages that most closely match the information content of the query. This is often referred to as *natural language understanding*. Also going in the other direction is considered NLP, i.e., generating or constructing human-language information from some computerized representation. The latter is often referred to as *natural language generation*. A unifying example is a machine translation system that includes both some type of language understanding and language generation components, together with a translation component, for translating a text phrase from one language into another.

2.1.1 Semantics

Semantics concerns the study of the meaning of natural language expressions and the relationships between them. In *computational semantics* the focus is on automatically constructing and reasoning with the meaning of natural language expressions. A common task in computational semantics is to calculate how similar, or related, linguistic items are based on their *semantic* meaning or content. We will refer to this as *semantic similarity assessment*. For instance, "pain" and "ache" have a rather high degree of semantic similarity since both refer to a type of painful sensation. This differs from, e.g., string similarity, where "pain" is a lot more similar to, lets say, "paint" than to "ache".

A *semantic similarity method/algorithm* usually relies on, and potentially produces, a representation that contains semantic similarity information in a way that enables the computer to reason with it, i.e., compute *semantic similarities*: *"A measure of semantic similarity takes as input two concepts, and returns a numeric score that quantifies how much they are alike."* (Pedersen et al. 2007). The utilized representation may be based on sets of (fixed) features that describes the concepts (see e.g., Smith and Medin (1981) Chapter 3), logical forms, graphs, or some type of combination. Nowadays feature sets are commonly treated as vectors of numeric elements, where each dimension represents a discrete feature, or where features are potentially distributed over multiple dimensions in a more

"sub-symbolic" representational manner (c.f. neural networks). Numeric vectors are convenient from a computer perspective in that it allow for the use of geometric algebra operations to, e.g., compute the likeness of vector pairs (see Section 2.1.3 about vector similarity). A vector-based representation is commonly referred to as a *vector space model* (VSM).

2.1.2 Language Processing Resources

Several approaches exist for manually developing lexical resources that model semantic similarity relations. These can be based on constructing rules (e.g., Appelt and Onyshkevych (1998)), thesauri (e.g., McCray et al. (1994)), ontologies (e.g., Bateman et al. (1995), Miller (1995)), and annotated data designed for machine learning (ML) algorithms (e.g., Pyysalo et al. (2007), Velupillai and Kvist (2012)). The more complex the task at hand is, the more manual labor is usually required in the development process. Thus, manually developed lexical resources tend to have very specific and restricted coverage in terms of what they represent, e.g., gene–protein relationships (Ashburner et al. 2000, Lord et al. 2003, Pyysalo et al. 2007).

By far the most comprehensive approach to modeling the terminology used in the medical domain is the development of the UMLS (NLM c) compendium. It consists of various lexical resources comprising primarily the vocabulary in medical research literature and clinical documentation, it also contains a mapping between the different vocabularies therein. SNOMED-CT (NLM b) represents medical terms in an ontological representation. SNOMED-CT originated as a resource for the English language, but has later been, or is currently being, translated into several other languages, primarily Spanish, Danish, Swedish, Dutch and French. MeSH (NLM a) is a thesaurus developed to index health research literature. It was originally made for English, but has later been translated or mapped to several other languages. ICD (the latest version being the 10th — ICD-10) (World Health Organization 1983) is a hierarchical medical classification, containing codes for diseases, signs and symptoms, etc., used primarily to classify diagnoses and treatments given to patients. Today the ICD classification has been translated into multiple languages.

The approach of manually developing lexical resources is well suited for modeling (semantic) relations on a conceptual level. However, with the vast information variety and complexity that natural language (free) text enables, it is very costly and challenging to conduct such modeling manually in a way that covers the language in its entirety. The same goes for enabling mappings between modeled concepts and the vocabulary — including the correct meaning of words and phrases as they are used in the text. An example illustrating some of the underlying challenges is

found in Suominen (2009), page 30, where it is reported that a single medicine has over 350 different spellings in clinical notes. This is one reason why normalization of the vocabulary in clinical notes has been the focus of several shared tasks, such as in the ShARe/CLEF eHealth Evaluation Lab 2013 (Suominen et al. 2013).

An alternative approach to manual lexical resource development is to model textual semantics in an automated and corpus-driven way. Methods in *distributional semantics* focus on learning/inducing semantic similarity from statistical information about word usage in a large corpus of unannotated text. In this thesis such a corpus is referred to as *training corpus* — typically being a collection of thousands or millions of unannotated documents. These methods are based on the *distributional hypothesis* (Harris 1954), which states that *linguistic items with similar distributions in language — in the sense that they co-occur with overlapping context — have similar meanings*. Two linguistic items, e.g., two words, having a similar "meaning" according to this hypothesis implies that they, statistically speaking, have been commonly used with the same or similar contextual features in the training corpus. For instance, they have co-occurred with the same neighboring words, or they have been often used within the same documents. The study of utilizing statistical approaches in computational semantics is sometimes referred to as *statistical semantics* (Weaver 1955, Furnas et al. 1983). The goal is to utilize statistical features in text to calculate a semantic similarity score between linguistic items that agrees with human judgement regarding the similarity of the items in a given context (c.f. "pain" and "ache").

Intuitively, relations between certain textual concept are difficult to obtain through purely statistical approaches, in particular those requiring complex implicit knowledge. For example, the relationships between known genes and proteins (Pyysalo et al. 2007). However, several hybrid approaches have been introduced that combine distributional information with lexical resources (Turney and Pantel 2010). For instance, Chute (1991) used Latent Semantic Analysis (LSA) to construct a distributional semantic model from the UMLS metathesaurus that enables matching of free-text inquiries with UMLS concepts; Henriksson et al. (2013b) constructed semantic models using Random Indexing, constructed from a corpus of clinical text, to extract synonyms for SNOMED-CT concepts/classes. Faruqui et al. (2015) performed retrofitting of word context vectors in various semantic models using lexical information from resources such as WordNet.

Distributional semantic methods have become popular due to their purely statistical approach to computational semantics and semantic similarity computation (Turney and Pantel 2010). Underlying factors are the increasing availability of large corpora and computational power. These methods enable rapid construction of new *semantic models* reflecting the semantic similarities in the languages and

domains in the utilized training corpus. Thus no costly manual labor is required for constructing annotated training data or lexical resources. As the training phase is "data-/corpus-driven", it is usually executed in a fully unsupervised manner. Also since the underlying training mechanisms are not dependent on the language of the training corpora, such methods can be classified as being language independent.

A training corpus used with distributional semantic methods consist commonly of only unannotated text that has been *pre-processed* to a certain degree. Pre-processing aims to improve the desired semantic representations in the resulting semantic model in various ways. *Tokenization* is, in its simplest form, about first splitting documents into sentences and finally into tokens or words/terms. We will be using 'terms' and 'words' rather indistinguishably, the main difference is that terms may contain multiple words (e.g., "car wheel" and "Yellowstone National Park"). Such multi-word terms and expressions can be recognized through a dictionary, rules, statistical co-occurrence information (*collocation segmentation*), annotated training corpora, or hybrid solutions (see e.g., Smadja (1993)). *Lemmatization* or *stemming* is a way to normalize a corpus by reducing the number of unique words. This is done by changing each word into their root form by removing and/or replacing words or suffixes (e.g., when using lemmatization "vocabularies" becomes "vocabulary", while with stemming "vocabularies" becomes "vocabulari"). Further, this tends to result in an increased distributional statistical basis for the remaining words since the vocabulary is reduced. The same also becomes a consequence of *lowercasing* words (e.g., "She" becomes "she"). Such normalization can be seen as a trade-off between precision and recall. E.g., lowercasing means you can no longer distinguish between proper nouns like "Apple" and common nouns like "apple". However, this will often improve recall since capitalized sentence-initials will not be confused with proper nouns. As distributional semantic models tend to emphasize high-frequent words and word co-occurrences, it is common to exclude "stop words" by filtering the corpus through a stop word list consisting of words that have little discriminative power in general or in a specific domain (e.g., "a" and "the"). *Part-of-speech tagging* and *dependency parsing* is something one can do to enrich the text with additional linguistic knowledge.

2.1.3 The Vector Space Representation

Vector spaces, or vector space models (VSMs), are by far the most common underlying representations in distributional semantics. VSMs were first introduced in text processing for the purpose of information retrieval by Salton et al. (1975). The underlying principle is to let each textual unit in a *training corpus*, such as words, sentences and documents, be represented as a multidimensional vector, or tensor. These vectors are referred to as *context vectors* (e.g., word context vector), representing the "contextual meaning" of the corresponding textual unit in

the utilized training corpus. A collection of these vectors constitutes the content of a vector space — a vector space model. Multidimensional vectors have the capacity to encode a lot of language information (e.g., semantics), where each element/dimension encodes a certain feature of the textual unit it represents. In many VSMs, particularly those whose dimensions have been compressed or reduced in some way (e.g., through some explicit dimension reduction operation or indirect feature prediction), these vectors' dimensions do not necessarily correspond to any known features of the language. Thus such vectors can be a composition of "sub-symbolic" features.

In addition to being an efficient way of representing textual information, VSMs allow for efficient ways of calculating similarities between vectors. To measure the similarity/dissimilarity between two vectors, it is common to use some type of algebraic distance function or metric. Different metrics tend to emphasize various types of semantic similarity (Lenci and Benotto 2012). In the work presented in this thesis the *cosine similarity metric* has been used. It calculates the cosine of the angle between two vectors. They can be called $\overrightarrow{x}$ and $\overrightarrow{y}$, thus turning the angle into a value between 0 and 1[2].

$$CosSim(\overrightarrow{x}, \overrightarrow{y}) = \frac{x \cdot y}{||x||\,||y||} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

The cosine similarity metric is often used because it is insensitive to the magnitude of vectors. When comparing two word context vectors using the cosine similarity metric, if their cosine similarity is 1 or close to 1, they have a high similarity. The opposite is the case if the cosine similarity is low. However, cosine similarity values in between 0 and 1 should only be interpreted relative to each other within a single VSM because it is often difficult to map these to any absolute truth outside a VSM. What these similarity values reflect according to a linguistic definition, if any, depends on how context vectors are constructed — what features they contain — and what similarity metric is used. A model may for example reflect that two words with a high cosine similarity value are similar because they are (near) synonyms, antonyms, metonyms, or morphological variants of the same word. It could also reflect that one or both words are abbreviations pointing to a concept that has the same or similar meanings, or that one or both are misspellings of the same or similar words. On the phrase/sentence level, it is common to use notions like paraphrasing and entailment, while topic and structural similarities may be used for document level similarities. Then again, such classifications may not be

[2]The cosine similarity metric can potentially return values between -1 and 1 if the vectors contain negative values.

of much importance in various tasks in computational semantics, where instead the main concern is the intra-model (semantic) similarities in relation to a specific task.

A common use-case for a VSM is to find how similar its constituent linguistic items, e.g., words or documents, are in relation to a given query. This is done by first retrieving, or possibly constructing, a vector representing the query, then compute cosine similarity between it and all other context vectors in the model. In this way we can rank the constituent context vectors according to their (semantic) similarity to the query. For example, if we have a VSM of word context vectors, and we query the model with the word "foot", we calculate the cosine similarity value between the context vector belonging to "foot" and all other context vectors in the model. This will give us a list of similarity values of each word relative to the query. By sorting this list (based on cosine similarity values) we can rank the words based on how similar they are to "foot", as illustrated together with the query word "pain" in Table 2.1. VSMs containing word context vectors are sometimes referred to as *word spaces* or *word space models* (WSMs).

foot	(jalka)	$CosSim$	pain	(kipu)	$CosSim$
lower limb	(alaraaja)	0.5905	pain sensation	(kiputuntemus)	0.5097
ankle	(nilkka)	0.3731	ache	(särky)	0.4835
limb	(raaja)	0.3454	pain symptom	(kipuoire)	0.4173
shin	(sääri)	0.3405	chest pain	(rintakipu)	0.4042
peripheral	(periferisia)	0.3112	dull pain	(jomotus)	0.4000
callus	(känsä)	0.3059	backpain	(selkäkipu)	0.3953
top of the foot	(jalkapöytä)	0.2909	pain seizure/attack	(kipukohtaus)	0.3904
upper limb	(yläraaja)	0.2879	pain status	(kiputila)	0.3685
peripheral	(perifer)	0.2875	abdominal pain	(vatsakipu)	0.3653
in lower limb	(alaraajassa)	0.2707	discomfort	(vaiva)	0.3614

Table 2.1: Top 10 most similar words to the query words "foot" and "pain", together with the corresponding cosine similarity scores. The results are derived from a distributional semantic model trained using W2V on a corpus of clinical text. The words have been translated from Finnish to English.

There are endless ways of generating context vectors in terms of what features define the semantic relations they capture and how these are weighted. A method introduced by Salton et al. (1975) works by deriving term-by-document statistics from a document corpus, generating a term-by-document matrix/model. The rows of the term-by-document matrix represent word context vectors, and the columns represent document context vectors. Here each dimension of a word context vector reflects how many times that word has occurred in each document, each dimension corresponding to one document. As an intuitive example, let us assume that we

have the following three short documents:

D1: The patient is suffering from an aching neck.

D2: The patient is experiencing pain in the neck.

D3: Take a taxi to the station.

By *preprocessing* these, through stemming, lowercasing and removal of stop words, they become:

D1: patient suffer ache neck

D2: patient experience pain neck

D3: take taxi station

Further, one can now create a term-by-document matrix based on the frequency of each word in each document, as illustrated in Table 2.2.

	D1	D2	D3
patient	1	1	0
suffer	1	0	0
ache	1	0	0
neck	1	1	0
experiencing	0	1	0
pain	0	1	0
take	0	0	1
taxi	0	0	1
station	0	0	1

Table 2.2: VSM example, term-by-document matrix, generated from three documents.

Statistically, words that occur in many of the same documents, i.e, occur in the same contexts, will have context vectors (rows) of high similarity to each other according to the cosine similarity metric. Likewise, documents containing many of the same words will have corresponding document context vectors (columns) of high similarity. Figure 2.1 illustrates on an intuitive level how similarities between the above documents (their context vectors) can be viewed according to vector angles (left); or as relative distances in a 2D semantic space (right). Such term-by-document models are particularly common in *information retrieval* (IR) (Manning et al. (2008), Chapter 6).

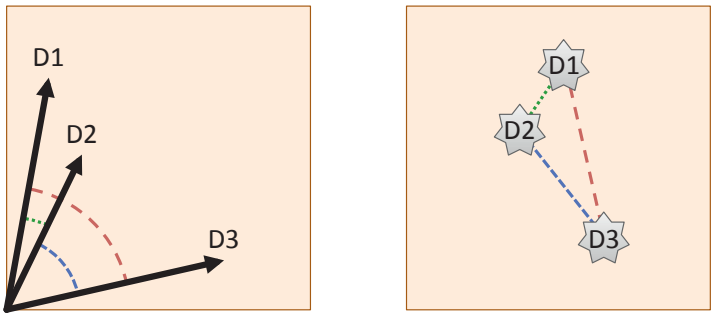

Figure 2.1: Intuitive illustration of how similarities between the three documents in Table 2.2 (their context vectors) can be viewed according to vector angles (left); or as relative distances in a 2D semantic space (right).

One approach to generating word context vectors is through constructing word-by-context models. Lund and Burgess (1996) use the neighboring words as context in their *hyperspace analogue to language* (HAL) method, which defines the semantic meaning of a word based on its neighboring words throughout a corpus. In this way, two words that co-occur with many of the same word neighbors, statistically throughout the training corpus, will have a high semantic similarity. HAL also applies a dimension reduction method to post-compress the matrix based on discarding the columns with lowest variance. Constructing a model from word co-occurrence information is typically done using the *sliding window* technique, where a window of a fixed size is slid across each sentence in the training corpus, iteratively updating each word based on the neighboring words. The size of the sliding window will naturally have an effect on the resulting semantic space, but the exact influence of this parameter seems to be task specific. For example, a window of size ten (5+5, left and right sides of the target word) has been shown to work well for modeling synonymy from clinical corpora (Henriksson et al. 2013a). The sliding-window approach is illustrated in Figure 2.2 where the size of the window is four (2+2). Table 2.3 shows how the resulting VSM becomes from the three example documents/sentences above.

Word context vectors, being the rows of a term-by-context matrix, or model, has what can be referred to as *second-order* co-occurrence relations between them since vector similarity is based on having similar neighbors. By measuring the cosine similarity between each word pairs, we can create a *similarity matrix* as in Table 2.4.

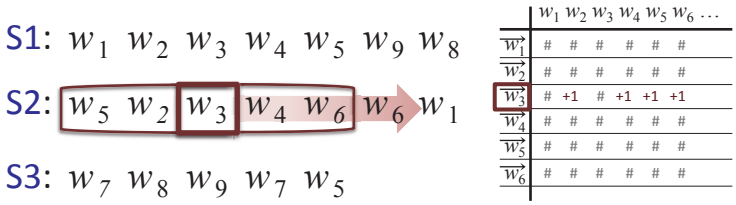

Figure 2.2: Illustrating training of a word-level co-occurrence distributional semantic similarity model using a "sliding window" with a size of four (2+2).

	patient	suffer	ache	neck	experience	pain	take	taxi	station
patient	0	1	1	0	1	1	0	0	0
suffer	1	0	1	1	0	0	0	0	0
ache	1	1	0	1	0	0	0	0	0
neck	0	1	1	0	1	1	0	0	0
experience	1	0	0	1	0	1	0	0	0
pain	1	0	0	1	1	0	0	0	0
take	0	0	0	0	0	0	0	1	1
taxi	0	0	0	0	0	0	1	0	1
station	0	0	0	0	0	0	1	1	0

Table 2.3: Word-by-context matrix, constructed using a sliding window with a size of four (2+2). Each row represents a word context vector.

	patient	suffer	ache	neck	experience	pain	take	taxi	station
patient	–	0.2887	0.2887	1.0	0.2887	0.2887	0.0	0.0	0.0
suffer	0.2887	–	0.6667	0.2887	0.6667	0.6667	0.0	0.0	0.0
ache	0.2887	0.6667	–	0.2887	0.6667	0.6667	0.0	0.0	0.0
neck	1.0	0.2887	0.2887	–	0.2887	0.2887	0.0	0.0	0.0
experience	0.2887	0.6667	0.6667	0.2887	–	0.6667	0.0	0.0	0.0
pain	0.2887	0.6667	0.6667	0.2887	0.6667	–	0.0	0.0	0.0
take	0.0	0.0	0.0	0.0	0.0	0.0	–	0.5	0.5
taxi	0.0	0.0	0.0	0.0	0.0	0.0	0.5	–	0.5
station	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	–

Table 2.4: Similarity matrix derived from the word context vectors in Table 2.3.

The similarity matrix in Table 2.4 can potentially be visualized as vectors or points in a semantic space, where their similarities are represented by their relative angles or distances, similarly to the illustration in Figure 2.1, with words instead of documents.

The semantic relations, represented in distributional semantic models, tend to be greatly influenced by common and frequent words occurring in many documents, words that often add little or nothing to the semantic meaning of a document. To counter this, one can re-weight the vector/matrix elements of a term-by-document matrix, using some weighting function. A common weighting method is to multiply the term/word frequencies by their corresponding inverse document frequency (TF*IDF) (Sparck Jones 1972).

$$tfidf\,(t, d, D) = freq(t, d) \times idf\,(t, D)$$

$$idf\,(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

Where:

- t is the term/word in question.
- d is a document.
- D are all documents in the corpus.
- N is the total document count, $|D|$.
- $|\{d \in D : t \in d\}|$ is the number of documents in which t occurs.

The purpose of TF*IDF weighting is to reduce the influence (weight) of words that occur in almost all documents and therefore have little value in discriminating one document from another. At the same time it increases the importance of words that are more rare and limited to a few documents, as these are potentially important to the topic of a document. TF*IDF weighted term-by-document matrices/models are used in various popular search engines, such as Apache Lucene (Cutting 1999).

As mentioned earlier, Latent Semantic Analysis (LSA) is a popular method for constructing (distributional) semantic models. LSA reduces the dimensionality of the VSM, while also having it emphasize latent correlations between words (and documents) through discovering higher order co-occurrence relations within a corpus (second order and above). Landauer and Dumais (1997) achieved human-level performance scores using LSA on the Test of English as a Foreign Language (TOEFL), a test where one has to choose the correct (closest) synonym among four alternatives for each query word. These scores have later been improved upon by others, through using LSA or other methods (see, e.g., Bullinaria

and Levy (2012)). Examples of more recent VSM-based distributional semantic methods are: Holographic Reduced Representations (HRR) (Plate 1991); Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999); Non-negative matrix vectorization (NMF) (Lee and Seung 1999); Random Indexing (RI) (Kanerva et al. 2000); Latent Dirichlet allocation (LDA) (Blei et al. 2003); various neural network-based language models, most popular being the Word2vec (W2V) implementation (Mikolov et al. 2013b). RI and W2V are used in various ways in the experiments in this thesis, and will be described in more detail in Sections 2.1.4 and 2.1.5.

2.1.4 Random Indexing

RI (Kanerva et al. 2000) is a method for building a *compressed* VSM with a fixed (reduced) dimensionality, and is done in an incremental fashion. This technique was originally intended as a way of overcoming the performance issues associated with LSA implementations at that time (computational complexity, scalability and memory requirements). Due to its computational efficiency, RI remains popular today for training on large corpora, such as MEDLINE/PubMed abstracts or articles (Cohen 2008, Jonnalagadda et al. 2012, Pyysalo et al. 2013), and social media (Sahlgren and Karlgren 2009). It has been shown to perform well, and is comparable to other methods, such as LSA, in a range of semantic similarity assessment tasks, including the TOEFL synonym test (Sahlgren and Swanberg 2000, Karlgren and Sahlgren 2001).

RI involves the following two steps:
Step 1 - Initialization: first, each word/term in the training corpus is assigned an *index vector* as its unique signature in the VSM. Index vectors have a predetermined dimensionality and consist mostly of zeros together with a small number of randomly distributed 1's and -1's — uniquely distributed for each unique word. This is based on the *Johnson Lindenstrauss Lemma* (Johnson and Lindenstrauss 1984), as discussed by Cohen et al. (2010), stating that distances between points in a high-dimensional space will be approximately preserved when projected into a lower-dimensional subspace. In other words, vectors being orthogonal in the high-dimensional space are assumed to be "near orthogonal" in the lower-dimensional subspace. Thus index vectors will have pairwise similarities, according to the cosine similarity metric, close to 0.
Step 2 - Training: the second step is the training step where context vectors are generated/induced for each unique word in the corpus. This is most commonly done using a sliding window of a fixed size (e.g., 2+2) to traverse the training corpus, inducing context vectors by superimposing the index vectors of the neighboring words in the window, as illustrated in Figure 2.3.

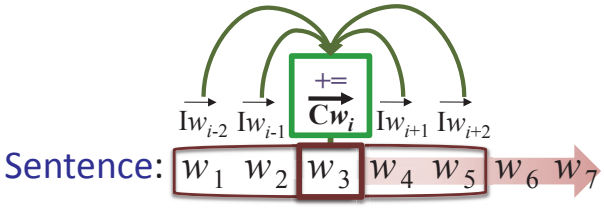

Figure 2.3: Illustrating how the training in RI works. The word context vector $\overrightarrow{\mathrm{C}w_i}$ is updated by adding the index vectors of its neighbors, i.e., through superimposing it with the neighbouring word index vectors $\overrightarrow{\mathrm{I}w_{i-2}}$, $\overrightarrow{\mathrm{I}w_{i-1}}$, $\overrightarrow{\mathrm{I}w_{i+1}}$ and $\overrightarrow{\mathrm{I}w_{i+2}}$.
This is a slightly modified version of Figure 3 in Moen et al. (2015), Paper C in this thesis.

As the dimensionality of the index vectors is fixed, the dimensionality of the vector space will not grow beyond the size $W \times Dim$, where W is the number of unique words in the vocabulary, and Dim being the pre-selected dimensionality to use for the index vectors, and ultimately the context vectors. As a result, RI models are significantly smaller than a full term-by-context model, which again make them a lot less computationally expensive in terms of storage and similarity computation. Additionally, the method is fully incremental in that additional training data can be added at any given time without having to retrain the model. It is also parallelizable and scalable, meaning that it allows for rapid training on very large corpora in a distributed on-line fashion. Using the JavaSDM implementation [3], with default parameters except a dimensionality of 800, the training on a Finnish clinical corpus (see Section 1.4.4) consisting of about 64 million words has an execution time of about 25 minutes. This is on a computer with the following hardware: Intel Core i7-3770 CPU @ 3.40GHz, 4 cores, 16GB RAM.

Some variants of the RI-based approach have been introduced, such as Random Permutations (RP) (Sahlgren et al. 2008) and Reflective Random Indexing (RRI) (Cohen et al. 2010), as well as cross-lingual variants (Sahlgren and Karlgren 2005).

2.1.5 Word2vec — Semantic Neural Network Models

The Word2vec (W2V) (Mikolov et al. 2013a) method/framework relies on using an artificial neural network to construct *neural network language models*. The models it constructs are vector-based and have been found to perform well in a range of semantic similarity assessment tasks (Baroni et al. 2014). Through training the network on a corpus, distributionally similar words are given similar vector

[3] `http://www.nada.kth.se/~xmartin/java/` (accessed 1st March 2016)

representations (i.e., context vectors).

W2V stems from the field *Deep Learning* (LeCun et al. 2015, Collobert et al. 2011). It uses a somewhat simplified neural network model, consisting of an *input layer* with as many input nodes as there are unique words (vocabulary items), a *hidden linear projection layer* with node count equal to the predefined dimensionality of the vector space, and finally a *hierarchical soft-max output layer* predicting the same words as the input layer (Morin and Bengio 2005, Mnih and Hinton 2009).

The context used for training is typically a sliding window. W2V has two training procedures/architectures: "continuous bag-of-words" (CBOW), and "continuous skip-gram model" (Skip-gram). The CBOW training approach aims to predict each word in the training corpus based on its context (co-occurring words). For each target word, the words corresponding to its context are activated in the input layer sequentially, i.e., the values of their corresponding input notes are set to 1 while the rest are 0. The expected/correct output for each training case is the correct target word in the output layer. Each target word and its connected weights are subsequently adjusted to decrease the error between the network outputs (normalized with soft-max) and the training cases using the *back-propagation* procedure (McClelland et al. 1986). This procedure is repeated for all training pairs, often in several passes over the entire training corpus, until the network converges and the error does not decrease any further. Now each word of the input layer has a context vector given by the set of weights connecting its corresponding input node to the hidden layer, as illustrated in Figure 2.4. The Skip-gram training approach predicts each individual context word (output layer) given the corresponding target word (input layer).

To understand on an intuitive level why the network learns efficient representations, i.e., distributional semantic models, we can consider the two-step process of the prediction: first, the input layer is used to activate the hidden, representation layer; and second, the hidden layer is used to activate the output layer and predict the context word. To maximize the performance on this task, the network is thus forced to assign similar hidden layer representations to words that tend to have similar contexts. Since these representations form the resulting model, distributionally similar words are given similar vector representations (c.f. context vectors).

One of the main practical advantages of the W2V method (CBOW/Skip-gram) lies in its relatively low complexity, giving it great scalability and allows for training on billions of words of input text in the matter of several hours. Using the optimized version in the Gensim implementation of Word2vec[4], with default pa-

[4] `https://radimrehurek.com/gensim/models/word2vec.html` (ac-

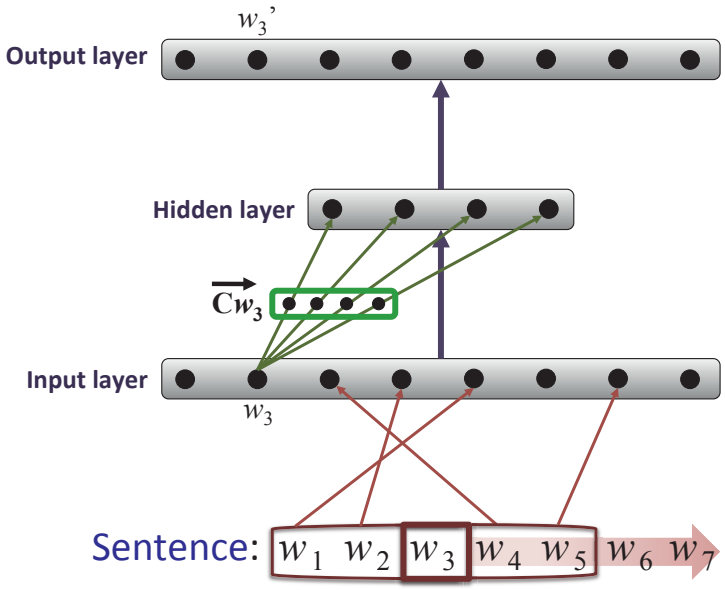

Figure 2.4: Illustration of how training happens in the W2V implementation of CBOW. A sliding window with the size of four (2+2) is moved over the text, word by word. The input layer nodes of the network corresponding to the words in the context window of the word w_3 are activated. The error in the output layer prediction and the expected prediction for the focus word w_3 is back-propagated through the network. When the training is completed, the context vector $\overrightarrow{Cw_3}$ constitutes the set of weights connecting the input layer node for w_3 and the hidden layer.
This is a slightly modified version of Figure 6 in Moen et al. (2015), Paper C in this thesis.

rameters except a dimensionality of 800, the training on a Finnish clinical corpus (see Section 1.4.4) consisting of about 64 million words has an execution time of about 20 minutes. This is on a computer with the following hardware: Intel Core i7-3770 CPU @ 3.40GHz, 4 cores, 16GB RAM. Shorter execution times are achieved when using a fully C-based implementation/package[5].

Neural network-based methods are based on *predicting* the target word or its context features. This differs from *count-based* methods such as RI and LSA that more directly *count* co-occurrences. Baroni et al. (2014) showed that prediction-based models, represented by W2V CBOW, achieved better results than a set of count-based ones in a range of tasks focusing on word-level semantic similarity

cessed 1st March 2016)

[5] `https://code.google.com/p/word2vec/` (accessed 1st March 2016)

assessment, including synonym detection/extraction in the TOEFL test and semantic similarity/relatedness classification. However, Levy et al. (2015) later showed that this performance advantage is likely due to smart parameter use and post-processing. An attractive property of W2V-based models is that they seem to preserve syntactic and semantic regularities (Mikolov et al. 2013c), e.g., $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ result in a vector similar to $\overrightarrow{queen}$. Levy and Goldberg (2014) revealed that these same regularities are also preserved, to the same extent, in count-based models when using some alternative similarity metric (i.e., not the cosine similarity metric).

2.1.6 Compositionality in Vector Space Models

So far this chapter has mainly discussed ways to construct distributed semantic models of words, representing word meaning by context vectors. However, in the experiments presented in this thesis, context vectors representing sentences, clinical notes, and care episodes have also been used. This is accomplished through performing some type of *compositionality* (Frege 1892, Montague and Thomason 1976) to ensemble such vectors from the constituent word context vectors. The idea is that a composition of word context vectors will result in a vector that captures the combined meaning of these words. Partee et al. (1990) explains *"The Principle of Compositionality"* as follows: *"... The meaning of a complex expression is a function of the meaning of its parts and of the syntactic rules by which they are combined."*.

Landauer et al. (1997) conclude that much information regarding the semantic similarity of texts, essays in this case, is carried by the semantic similarity between constituent words independently of their order. Thus one straightforward approach to representing multi-word items in VSMs is to treat a collection of words, e.g., a sentence, as simply a "Bag of Words" (BoW), where their order is irrelevant. In this approach, a composed vector, e.g., a sentence context vector, is generated through simply pointwise summing its constituent word context vectors (also referred to as superimposing). In addition to vector *addition*, other alternatives includes vector *multiplication* and the use of *circular convolution* in *holographic reduced representations* (Plate 1991). To reduce the influence of the magnitude of each individual vector, it is common to first normalize vectors to unit length. Further re-weighting can be done by applying TF*IDF weighting (see Section 2.1.3).

There are also ways to compose vectors that incorporate some information about the constituent word order. Such approaches typically focus on constructing context vectors for n-grams, short phrases or larger linguistic items that in some way emphasize the order of the constituent words (see, e.g., Guevara (2011), Mikolov et al. (2013b), Le and Mikolov (2014)). Such a VSM may, as an example, contain

vectors representing the phrases “dog eats rabbit”, “rabbit eats dog”, and “dog eats dinner”. Here the phrase “dog eats rabbit” should intuitively be closer to “dog eats dinner” than “rabbit eats dog” in the semantic space. Another way to retain word order information when performing similarity assessment is to simply view each sentence, document, etc., as a list of their constituent word context vectors. In this approach the order information is not actually modeled into the semantic model, but is calculated at retrieval time. When computing the similarity of two sentences one may apply some sequence aligning algorithm, e.g., *Needleman-Wunsch* (Needleman and Wunsch 1970), to compute a total similarity score based on word alignment and word pairwise cosine similarity scores (see, e.g., Feng et al. (2008)).

2.2 Distributional Semantic Models and Clinical Language Processing

Clinical language in this thesis is defined as the language clinicians use when documenting patient care, mainly in the form of written text notes stored in the patients’ health records. As the focus is on clinical text written in hospitals, we refer to physicians/doctors and nurses involved in clinical care in hospitals as *clinicians*. There are often physicians with different medical specializations, located at different wards within the hospital, involved in the treatment during a care episode, such as internal medicine, cardiology and surgery. Thus the clinical notes, or narratives, that they write tend to reflect the different tasks being performed at the respective wards. In this thesis the term ‘clinical note’ refers to any of the different notes that the various specialists write to document patient care. During a care episode, a sequence of clinical notes are written (as illustrated in Figure 1.1). These are stored in the patient’s health record, which again is stored digitally in an electronic health record (EHR) system.

Clinical notes contain highly domain-specific terminology (Rector 1999, Friedman et al. 2002, Allvin et al. 2010), and clinical language can be regarded as a *scientific sub-language* (Meystre et al. 2008). Some features of the written clinical language/notes are:

- The different professions and individual clinicians tend to have their own way of documenting — documenting observations, symptoms, diagnoses, and their reasoning and speculations.
- Each note may have a different author, including those belonging to the same care episode.
- The texts usually contain ungrammatical language, incomplete sentences,

abbreviations, and medical jargon.

- Authors do not necessarily utilize any common note structure.
- The written information tend to be highly domain- and case-specific, including a fair share of implicit information.
- All notes in one care episode are related to the care and treatment given to the same patient, meaning that they all are linked to a series of related events and often contain repeated and overlapping information. This is also the case when looking at the full health record belonging to a patient, since some information from one care episode could be relevant to a later one.

Figure 2.5 shows an example of a clinical note.

English translation:
61-years old female with Crohn's disease. Attended cycling event in Salo, flu prirorlry. Arfter cycling, experienced breathing difficulties and went to the emergency department and elevated herart enzymes and incompensation were found. Was admitted to the ICU for care of incompensation and pneumonia. In UKG 2.6. ef 30%. In coronary angiography, significrant stenoses in RCA, LCX and mrarin. Trordray, elective quadrurple bypass LITA-LAD, Ao-LOM-LPL and Ao-RBD, in which goord flow. Pre.op. left ventricle, the posteriror myocardium and septum contract lamely, ef about 35%, mitral valve 1-2/4 leak. Aortic cross-clamp time 1 h 32 min. Post.op. ef over 40%. On basis of the UKG-finding pre.op. Simdax-infusion was initiated. On arrival to ICU, haemodynamics was stable, norepinephrine administered. Cardiac index 3,2. Warming-up and weaning in ventilator.

Finnish original:
61-vuotias nainen jolla Crohnin tauti. Salossa ollessaan osallistunut pyöräilytapahtumaan, edelträvrästi flunssaa. Pyöräilyn jrälkeen hengitysvaikeuksien takia TYKSin ensiapuun ja todettu sydränentsyymit kohonneiksi ja inkompensaatiota. Otettu teho-osastolle inkompensaation ja pneumonian hoitoon. UKG:ssa 2.6. ef 30%. Koronaariangiossa merkitträvrät stenoosit RCA:ssa, LCX:ssa ja prärärungossa. Tränrärän elektiivinen neljrän suonen ohitus LITA-LAD, Ao-LOM-LPL ja Ao-RBD, joihin hyvrät virtaukset. Pre.op. vasemman kammion takaseinrä ja septum supistuvat vaisusti, ef noin 35%, mitraaliläpässä 1-2/4 vuoto. Aortan sulkuaika 1 t 32 min. Post.op. ef yli 40%. UKG-löydöksen perusteella potilaalle aloitettu jo pre.op. Simdax-infuusio. Teho-osastolle saapuessa hemodynamiikka stabiilia, noradrenaliini menossa. Cardiac index 3,2. Lämmitys ja vieroitus respiraattorissa.

Figure 2.5: Example of a clinical note. This is a fake case originally created in Finnish by domain experts, then translated into English. Common misspellings are included intentionally.
This is the same example as in Figure 2 in Moen et al. (2016), Paper E in this thesis.

Developing methods and systems for clinical NLP is hard for a number of reasons. Some of the main challenges are: lack of data available to software developers

and researchers, primarily due to the sensitivity of clinical information/text; lack of existing and robust text processing resources that support the broad range of (sub-) languages and applications, such as text parsers, computerized thesauri and ontologies; the cost of developing new or customized methods and resources for processing the text; the issues related to integrating NLP software into the clinical practice through existing and new information systems (Chapman 2010, Rector 1999, Kate 2012, Friedman et al. 2013, Meystre et al. 2008, Grabar et al. 2009, Kvist et al. 2011).

NLP has been applied to clinical text for a variety of tasks. Some examples are automatic event detection in health records (Mendonça et al. 2005), automatic concept indexing (Berman 2004), medication support (Xu et al. 2010), decision support (Demner-Fushman et al. 2009, Velupillai and Kvist 2012), query-based search (Grabar et al. 2009)) and automated summarization (Pivovarov and Elhadad 2015). Openly available tools designed for clinical NLP typically dependent on specialized and extensive knowledge resources to classify and reason with concepts in the text. Such resources that are commonly built in a manual fashion (see Section 2.1.2 for more information). Further, due to the domain specificity of clinical text, generic resources for computational semantics tend to be of limited use. However, the use of distributional semantic methods in this domain is promising due to their focus on learning semantic relations directly from unannotated corpora. This enables acquisition of semantic resources in an resource-lean manner.

Distributional semantic methods utilize statistics that are derived from the corpus used for training, thus the resulting models and its constituent context vectors will reflect the semantic similarity relations that are found in the utilized training corpus, which again reflects the domain of the corpus. Koopman et al. (2012) show that the domain-specificity of the corpora used for training such distributional semantic models is important for the content and quality of the resulting model with respect to the intended task. Pedersen et al. (2007) explore a set of measures for automatically judging semantic similarity and relatedness among medical concept pairs whose semantic similarity have been pre-assessed by human experts. These range from various measures based on lexical resources (WordNet, SNOMED-CT, UMLS, Mayo Clinic Thesaurus) to one based on a distributional semantic model trained on a corpus of unannotated clinical text. The latter measure uses the LSA method, and the semantic similarity between concept pairs is calculated using the cosine similarity measure applied to *concept context vectors* — assembled from the corresponding words. Pedersen et al. (2007) find that this measure performed at least as well as any of the other measures. Related work has also shown that distributional semantic models, induced automatically from large corpora of clinical text, or other types of medical text, are well suited as a fast and cost-efficient

approach to capturing and representing domain-specific terminology (Koopman et al. 2012, Cohen and Widdows 2009, Cohen et al. 2014, De Vine et al. 2014).

As an example, a semantic model trained with W2V CBOW on a fairly large corpus of clinical free-text notes is able to detect that the words "pain" and "discomfort" have a similar meaning or contextual use because they have a relatively high cosine similarity value — relative to the other words in the model. However, if some other type of corpus was used for training, e.g., a collection of newspaper articles, the resulting model would not necessarily contain the same semantic relations. The same goes for other domain-specific terms that clinicians use when documenting care. A number of these would not even be present in a newspaper corpus, let alone abbreviations and spelling mistakes that are common in clinical text. Paper A explores various combinations of distributional semantic models, trained on one of two different corpora — one containing clinical text and the other medical research literature — and evaluates these on the tasks of automatic extraction of synonyms and abbreviation-expansion pairs.

Karlgren and Sahlgren (2001) argue that text models and methods for constructing/training them still have a way to go in terms of capturing the actual language use, rather than the language in abstract. Most distributional semantic methods construct word space models that contain one context vector per unique word. This means that each word will have one semantic meaning, representing its "prototypicality", relative to the others in the semantic model, accumulated from all its occurrences with the utilized training features, e.g., neighboring words, in the training corpus. However, in reality the meaning of a word may vary greatly based on the context of its use, thus each word could potentially have multiple meanings or senses. Such words are referred to as polysemes or homonyms (Panman 1982). As an example, the word "discomfort" may refer to some type of physical pain, or it may refer to psychological/social inconvenience. Another example is the word "rock", which may refer to a type of music or a material. One direction in distributional semantics concerns training vector spaces that allow words to potentially have more than one representation, i.e., multiple context vectors (Reisinger and Mooney 2010, Schütze 1998, Neelakantan et al. 2014). This would intuitively be beneficial in a range of semantic similarity assessment tasks, including tasks in the clinical domain. Arguably these type of methods may enhance the information complexity represented in the resulting models, as well as their discriminative capabilities — discrimination between different local meanings of words found in the training corpus. This may further enhance the task and domain specificity of semantic models.

Having potentially multiple context vectors for each word from the training corpus means that a large(r) number of vectors have to be stored in the computer mem-

ory, in particular during training. In the approaches by Reisinger and Mooney (2010) and Schütze (1998) every contextual occurrence of each word throughout the training corpus is stored in memory before applying some type of clustering. Paper B explores a novel "multi-sense", or "multi-prototype", distributional semantic method that performs incremental clustering as part of the training phase. It builds on the RI method and retains the properties of RI concerning reduced dimensionality and on-line training. We evaluate this method at a semantic textual similarity task (Agirre et al. 2013), where the goal is to automatically assess and classify similarities between sentence pairs.

In Friedman et al. (2013) it is suggested that future work in clinical NLP should aim to exploit existing knowledge bases about medications, treatments, diseases, symptoms, and care plans, despite these not having been explicitly built for the purpose of clinical NLP. One way to potentially improve the task- and domain-specificity of distributional semantic models is to exploit more domain specific features of the training corpus for constriction. This may assist in forming a semantic space that better reflects the semantic relations of interest. Paper C explores the use of ICD-10-code labels (see Section 2.1) as training features in an attempt to induce the underlying relations into a distributional semantic model. This is used in a set of experiments that explores various ways of constructing distributional semantic models for the task of *care episode retrieval*, using only the free-text information in clinical notes for the retrieval process.

2.3 Automatic Text Summarization of Clinical Text

(Jones 1999) presents *factors* that one has to take into account in order to make a summarization system achieve its task. The three main categories of factors described are *input*, *purpose* and *output*. These factors have also been discussed and elaborated upon by (Hahn and Mani 2000, Afantenos et al. 2005). Following are the factors and their underlying properties and recommendations that we have identified as being most suited relative to the research goal (c.f. RG1)[6]. For further details about these factors, please see the mentioned papers.

Input Factors

- *Single document or multiple documents*: As the task is to summarize clinical free text notes written during care episodes, the system input would primarily be multiple documents — sequences of multiple clinical notes constituting care episodes — one care episode per summary that is to be produced.

[6]These properties and recommendations are the result of a literature study conducted relatively early on in the PhD period, but is currently unpublished material.

- *Structure*: Information about the structure of the document or documents can help in classifying the content. As each clinical note are parts of a continuous patient story, the approach should have a scope that covers the full care episode when assessing what the most relevant information is. If a predictable document/note structure is used by clinicians, this should be exploited.

- *Language*: The language specificity of the system is commonly determined by the underlying resources and tools that it relies on. Further, a knowledge-poor approach would potentially enable easy adaptation to a broad range of languages and sub-languages at a low cost.

Purpose Factors

- *Indicative, informative, and/or critical*: We strive towards a system that is able to provide an indicative overview of the free text documented for care episodes. Together with structured data (such as laboratory test results, images, diagnostic codes and personal information) it could help clinicians to quickly familiarize themselves with the content of individual care episodes and the patients problems, which is particularly useful if such information is needed urgently.

- *Generic or user-oriented*: The system should be both generic and user-oriented in order to meet the specialized information needs of clinicians. However, for (automated) evaluation purposes, we believe that producing generic summaries is the first thing to aim for.

Output Factors

- *Extracts or abstracts*: We aim for a extraction-based summarization approach, in which the summary is generated by selecting a subset of sentences from the relevant text. This approach is viable because a sizeable portion of clinical text summaries, such as discharge summaries, are created by copying or deriving information from clinical notes (Van Vleck et al. 2007, Sørby and Nytrø 2005, Meng et al. 2005, Wrenn et al. 2010).

- *Available domain knowledge*: It is common to distinguish between "knowledge-rich" and "knowledge-poor" systems based on the availability of data and domain knowledge resources for the system to exploit. As already mentioned, in particular for small (clinical) languages, few such specialized resources exists.

- *Output format*: The output should, at least as an initial approach, have a format that is similar to the notes that clinicians produce themselves the care process to enable automated evaluation against existing summaries (c.f. gold standard).

- *Quality (evaluation)*: The quality of a summarization system is commonly measured by its content-selection capability, presented as its output. Using manually created summaries — a so called *gold standard* — for comparison is a common way to evaluate the quality of a summarization system. However, creating manual summaries is a expensive and time-consuming process. We suggest exploring summaries constructed/written by clinicians during the care process for this purpose (see Paper D and E).

The most central issue in text summarization is to determine what information to include in a summary. In *extraction-based summarization* this concerns selecting a subset of sentences from the text that is to be summarized. Common techniques here are: Topic-based extraction (see, e.g., Carbonell and Goldstein (1998), Goldstein et al. (2000), Steinberger and Křišt'an (2007)), where relevance scores for sentences are computed with respect to one or more topics of interest; Centrality-based extraction (Patil and Brazdil 2007, Chatterjee and Mohan 2007, Erkan and Radev 2004, Mihalcea and Tarau 2004), where typically some underlying graph-based representation is used to calculate sentence significance based on the document coverage of the sentences relative to the other sentences. An important sub-task when applying these techniques is to avoid redundant information in the produced summaries. For this purpose it is common to apply some ways of checking for textual similarity overlap based on the Maximal Marginal Relevance (MMR) criterion (Carbonell and Goldstein 1998) or similar techniques. Distributional semantic models, in various forms, have been quite extensively used in the field of text summarization, see e.g., Luhn (1958), Chatterjee and Mohan (2007), Hassel and Sjöbergh (2007), Nenkova and McKeown (2011).

Several pieces of work have been identified focusing on the task of automatically generating summaries from the text in clinical notes. Liu (2009) uses the MEAD summarization toolkit. Van Vleck et al. (2007) perform structured interviews to identify and classify phrases that clinicians considered relevant to explaining the patient's history. Meng et al. (2005) use an annotated training corpus together with tailored semantic patterns to determine what information that should be repeated in a new clinical note or summary. Velupillai and Kvist (2012) focus on recognizing diagnostic statements in clinical text, learned from an annotated training corpus, and further to classify these based on the level of certainty. Extracted diagnostic statements are then used to produce a text summary. Bashyam et al. (2009),

Hirsch et al. (2015) reports the work on extensive clinical summarization systems. These apply various information extraction tools and resources to identify, classify and reason with entities and information in free text. Visualization is also an important part of these systems, including timeline-based visualization in Hirsch et al. (2015). Others have worked on more conceptual models for understanding and supporting generation of information summaries in the clinical domain (Sarkar et al. 2011, Abulkhair et al. 2013). However, to the best of my knowledge, summarization of clinical free-text information has been pursued by relatively few researchers. This is not surprising given the challenges related to clinical NLP and the task. This is also a prominent issue considering recent reviews and related work (Mishra et al. 2014, Pivovarov and Elhadad 2015, Kvist et al. 2011).

In extractive multi-document summarization there is a need to have the computer "understand" the *terminology* of, or semantic similarities between, the candidate sentences to determine if some information is repeated, redundant, or similar to some topic or query (Ferreira et al. 2016). For this task, distributional semantic models are commonly used. In Paper E we explore various distributional semantic models at the task of summarizing clinical notes for individual care episodes. We focus on exploring a resource-light approach that circumvents the need for manually developed knowledge and training resources tailored for the task.

Computer generated summaries are typically evaluated by comparing the summary with a *gold standard*, being one or more reference summaries that has been constructed manually, often in relation to a shared task[7]. To perform this comparison in an automated fashion, a computerized similarity metric is used. The *ROUGE evaluation package* (Recall Oriented Understudy for Gisting Evaluation) (Lin 2004) evaluates text similarity based on N-gram overlap. ROUGE metrics are commonly used in text summarization evaluation because its scores have shown to correlate well with human judgements (Lin 2004). Liu (2009) performs automatic evaluation of computer generated summaries of clinical notes by using the original discharge reports as gold summaries. An alternative type of evaluation is to do the content assessment manually. Lissauer et al. (1991) evaluate computer generated discharge summaries from neonate's reports by analysing if they contain the required information according to a guideline. In Papers D and E we apply the ROUGE evaluation package for evaluation of automatically generated clinical free-text summaries where the original discharge summaries are used as gold standard. Evaluation scores are then compared to manual evaluation, conducted in a

[7] A 'shared task' is here defined as a specific task proposed and organized by a dedicated committee that provide the necessary data and evaluation setup to the participants. A shared task is typically held in relation to a conference where there are multiple participating research groups who are competing to achieve the best results.

similar fashion as in the work by Lissauer et al. (1991).

Chapter 3

Paper Summaries

The first experiments focused on automated assessment of word-level similarities, which resulted in Paper A. More precisely this study concerned automatic detection of synonymic relations between words, including between full form words and their abbreviations. One motivation behind these experiments was to get an insight into what methods and parameters that produces models that best captures synonymic relations on a word level. Further, it is likely that a similar setup would also apply to sentence-level semantics — a central part of (sentence-level) extraction-based text summarization (c.f. Paper E). The next set of experiments, presented in Paper B, focused on sentence-level semantic textual similarity assessment. Here a method that performs automatic word-sense discrimination was evaluated. In relation to the work on automatic text summarization of clinical free-text notes, I wanted a way to retrieve care episodes that are similar to a target care episode, belonging to other (patients') hospital stays. This resulted in the work on care episode retrieval, presented in Paper C. The last set of experiments is related to automatic generation of summaries from clinical free text, from one care episode at a time, and evaluation of the generated summaries. This work is presented in Papers D and E. The work in Paper E is to a large extend based on the lessons learned from previous papers/experiments and includes many of the approaches, methods and models therein.

Below is an overview of the five papers in this thesis. For each paper there is: a *summary* of the main content; followed by a *retrospective view* discussing the work from a possibly more enlightened perspective.

3.1 Paper A: Synonym extraction and abbreviation expansion with ensembles of semantic spaces

Authors: *Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius and Martin Duneld*

3.1.1 Summary

Terminologies that account for variation in language use by linking synonyms and abbreviations to their corresponding concepts are important for enabling automated semantic similarity assessment and high-quality information extraction from clinical texts. Due to the use of specialized sub-languages in the clinical domain, manual construction of semantic resources that accurately reflect language use is both costly and challenging.

In this paper we explore the use of distributional semantic models for automated extraction of synonymic relations from clinical/medical text, including abbreviations (abbreviations to long forms and long forms to abbreviations). Models are trained on one or both of two corpora, one corpus consisting of Swedish clinical text and another consisting of Swedish medical journal articles. Two approaches to constructing the models are used, classic Random Indexing (RI) and the Random Permutations (RP) variant. Various model training parameters and ways of combining the retrieved candidate words/synonyms are explored.

Evaluation is done using two gold standards. For the synonym extraction task, MeSH terms and associated synonyms are used. And for the other two tasks (abbreviations to long forms and long forms to abbreviations), we used a list of medical abbreviation-expansion pairs. Only single-word terms were used for evaluation. For each query a list of ten candidate words are retrieved by the model(s) being evaluated. The results are measured primarily as recall among these ten (R, top 10), calculated from the proportion of expected candidate words that are among these. When combining the results from two or more models, their retrieved lists of scored candidate words are combined through summing or averaging over matching words.

We also explore the use of some post-processing filtering. For abbreviation-expansion extraction, the filtering is based on word-length threshold filtering and on checking for overlapping letters and their matching order. For synonym extraction, the filtering rule checks if the retrieved candidates has a cosine similarity above a given thresholds, together with checking if their rank (in the retrieved list) are above a given threshold. Also different word frequency thresholds are explored, i.e. words below a given threshold are removed from the gold standards.

We found that a combination of the two models (RI + RP), trained on a single corpus outperforms the use of one model in isolation. Furthermore, combining semantic models induced from the two different types of corpora further improved the results ($RI_{clinical}$ + $RI_{medical}$ + $RP_{clinical}$ + $RP_{medical}$), also outperforming the use of a conjoint corpus ($RI_{clinical+medical}$ + $RP_{clinical+medical}$). A combination strategy that simply sums the cosine similarity scores of candidate words — retrieved from each model — is generally the best performing one. Finally, applying the post-processing filtering rules yielded substantial performance gains on the tasks of extracting abbreviation-expansion pairs, but this is not the case for synonym extraction. A word frequency threshold in the range of 30-50 seems to be optimal. Best results achieved R = 0.39 for abbreviations to long forms ($R_{baseline}$ = 0.23), R = 0.33 for long forms to abbreviations ($R_{baseline}$ = 0.24), and R = 0.47 for synonyms ($R_{baseline}$ = 0.39).

This study demonstrates that ensembles of semantic models can yield improved performance on the tasks of automatically extracting synonyms and abbreviation-expansion pairs — improvements compared to using a single model. Further, this encourages further exploration in utilizing and combining different semantic models, trained with different parameters and context features, and/or trained on different types of corpora. This also includes exploring different ways of combining the model outputs during the retrieval phase. The methods, models and model combinations in this study could potentially be used in (semi-) automated terminology development in the clinical/medical domain, as well as in a range of other NLP tasks.

3.1.2 Retrospective View and Results/Contributions

This study gave valuable directions and insight into how to generate semantic models from clinical/medical free text that capture word-level similarities, reflecting similarity in terms of having the same or closely related synonymic meaning. Experienced gained here were important for further work on distributional semantic similarity models.

From a retrospective view, it would have been interesting to see how these methods, RI and RP, fare at the given tasks in comparison to, or in combination with, other distributed semantic methods/models such as LSA and more recent neural network-based methods, such as W2V CBOW and Skip-gram. Fundamentally different methods that does not rely on distributional semantics, such as more rule-based methods (e.g., Ao and Takagi (2005)), are likely to perform well when it comes to detecting relations between terms and their abbreviations.

3.2 Paper B: Towards Dynamic Word Sense Discrimination with Random Indexing

Authors: *Hans Moen, Erwin Marsi and Björn Gambäck*

3.2.1 Summary

Most distributional models of word similarity represent a word type by a single vector of contextual features, even though words often have more than one lexical sense (Reisinger and Mooney 2010, Huang et al. 2012). In this paper we present a novel method for learning and constructing a distributional semantic model that may contain more than one context vector, or "sense vector", for each unique word in the utilized training corpus. A common way of capturing multiple senses per word with the distributional semantic approach is to first construct and store one vector for each occurrence of a word in the training corpus — storing the features of each single word use. Then these vectors may be clustered in some way to create sense vectors. However, storing and clustering these vectors can be expensive as it generates a set of vectors equal to the word count of the training corpus. As an alternative, we introduce *Multi-Sense Random Indexing*, that performs on-the-fly incremental clustering of word senses, allowing multiple senses per unique word in the training corpus. A range of different measures for sentence similarity are explored that focus primarily on deriving these from the maximum bipartite similarities between the underlying words and their different senses. Various measures for word-sense alignment are illustrated in Figure 3.1.

For training the semantic models we use the CLEF 2004–2008 English corpus (CLE 2004). We use the STS 2012 and STS 2013 shared tasks' evaluation data (Agirre et al. 2012; 2013) for sentence similarity assessment, where the task is to score the similarity between sentence pairs with a number between 0 and 5. A *support vector regressor*[1] is trained/optimized using the training data accompanying the STS task for the purpose of mapping the cosine similarity scores to a final score between 0 and 5. The various multi-sense based performance scores are compared to those from using a classic (single sense) RI model. Performance scores are calculated using the mean Pearson product-moment correlation coefficient (PPMCC) (Lehman 2005), same as in the mentioned STS tasks.

Our experimental results did not show a clear systematic difference between single-prototype and multi-prototype models. The highest scores were achieved on the STS 2013 evaluation data, with a mean PPMCC = 0.46 with the multi-sense Hungarian Algorithm-based similarity measure, compared to a mean PPMCC = 0.45

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html` (accessed 1st March 2016)

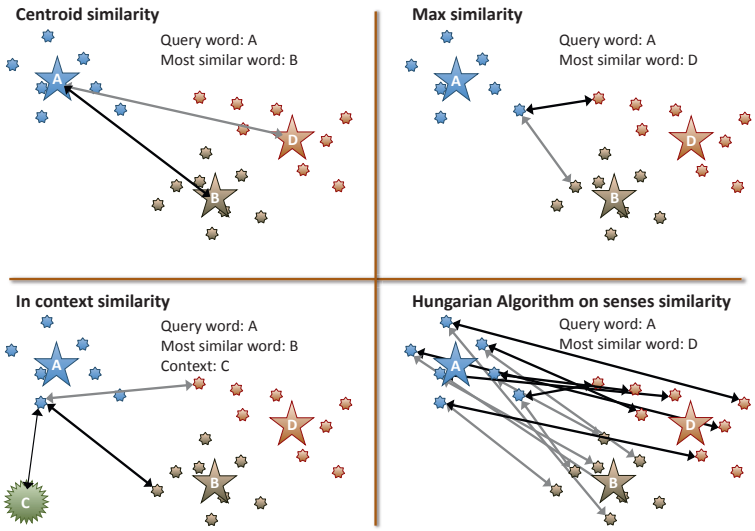

Figure 3.1: Various similarity measures tested with the multi-sense vector space model. In this 2D illustration the relative distances between words and senses reflect how similar they are. Large stars represent the centroid location of words, and the small stars represent their underlying senses.

achieved with single-sense Hungarian Algorithm-based similarity measure (see Kuhn (1955) for more information on the Hungarian Algorithm).

3.2.2 Retrospective View and Results/Contributions

The motivation for the multi-sense method introduced in this paper was to see if we could better capture the meaning of words by creating semantic models that learn potentially more than one context vector per word, i.e., "sense". At the same time we wanted to retain the incremental and compressed dimensionality features of the RI method. When calculating the similarity between two sentences, the aim was to select the most appropriate sense vector for each word based on a) the context defined by the other words in the same sentence, or b) the words (and their senses) in the other sentence. Then we calculated sentence similarity from word-pairs using the Hungarian Algorithm for word alignment, or composed two sentence context vectors before calculating their similarity. The results from the latter approach was not included in this paper since the Hungarian Algorithm-based approach generally performed better at the task. However, results using

such sentence context vectors with TF*IDF weighting have been reported in Marsi et al. (2013).

Table 3 in Marsi et al. (2013) shows that the top three strongest single similarity features, individually trained using a support vector regressor (Vapnik et al. 1997), are those based on character n-gram overlap. Although the approach rely on manually classified training data for the regressor, it is worth noting that this rather simple approach performs well when compared to the more complex ones.

This study, together with that presented in Marsi et al. (2013), provided us with valuable insight into sentence similarity assessment. This was also an opportunity to compare a variety of different approaches, including some not relying on distributional semantic models. In the experiments conducted in Marsi et al. (2013), a support vector regressor was used to learn a function that, for each sentence pair, take a range of different sentence similarity features as input, and the output is a single similarity score that is optimized to reflect some given training instances (manually scored with values in the range 0 to 5). Here it became clear that individual similarity features, also those achieving relatively weak scores individually, would contribute to increasing the final score when used in combination with others. Further, the lessons learned here were important to the sentence similarity calculations and clustering used in the summarization task presented in Paper E.

Despite that the results presented in this paper did not show systematic improvements over the evaluation scores gained by *not* using the multi-sense method, this approach calls for further research. The more recent publication by Neelakantan et al. (2014), who applied a similar training algorithm as ours in their multiple-sense (W2V) Skip-gram-based method, indicates that this direction in distributional semantics has a certain actuality. Also, after publishing our paper, we did more exploration with various parameters, and were able to improve the mean PPMCC of the multi-sense Hungarian Algorithm-based similarity measure to 0.49, up from 0.46, on the STS 2013 evaluation data. In the future, it would also be interesting to evaluate this method on other tasks than sentence-level similarity assessment.

There are still many unresolved and open questions regarding parameters and training features to use during training and how to do the word and sense extraction/retrieval and similarity calculations. Finally, if such an approach is able to improve upon the existing state-of-the-art in automated sentence similarity assessment, there is little doubt that it should also have a positive influence on tasks such as automatic terminology development and (clinical) text summarization.

3.3 Paper C: Care Episode Retrieval: Distributional Semantic Models for Information Retrieval in the Clinical Domain

Authors: *Hans Moen, Filip Ginter, Erwin Marsi, Laura-Maria Peltonen, Tapio Salakoski and Sanna Salanterä*

3.3.1 Summary

Electronic health records (EHRs) are used throughout the health care sector by professionals, administrators and patients, primarily for clinical purposes, but they are also used for secondary purposes such as decision support and research. The vast amounts of information in EHR systems complicate information management and increase the risk of information overload. Therefore, clinicians and researchers need new tools to manage the information stored in the EHRs. A common use case is, given a — possibly unfinished — care episode, to retrieve the most similar care episodes among the records.

This paper presents several methods for information retrieval, focusing on the task of *care episode retrieval*. The experimental setup is illustrated in Figure 3.2. Care episode similarity is calculated based on their textual content. This is achieved through constructing different distributional semantics models from a corpus of clinical text, and then applying the cosine similarity measure. Methods used to construct these models include variants of RI and W2V. A novel model construction approach is introduced that utilize the ICD-10 codes attached to care episodes as training features to better induce domain-specificity in the resulting distributional semantic model. When calculating the similarity between care episode pairs, we explore a set of different approaches for aligning and comparing them in terms of their underlying clinical notes.

We report on experimental evaluation of care episode retrieval that circumvents the lack of human judgements regarding episode relevance. Results are reported as: precision among the top-10 retrieved care episodes (P@10); precision at the R-th position in the results (Rprec), where R is the number of correct entries in the gold standard; mean of the average precision over all queries (MAP). The results suggest that several of the proposed methods outperform a state-of-the-art search engine (Lucene) on the retrieval task. The best results were achieved when using the ICD-10-based semantic model, constructed using a modification of the W2V Skip-gram algorithm (W2V-ICD), and when treating care episodes as single conjoint text documents (i.e., not as a series individual clinical notes). On this setup, the best performing method, W2V-ICD, achieved: MAP = 0.2666, P@10 = 0.3975, Rprec = 0.2874. In comparison, on the same setup, Lucene achieved: MAP = 0.1210, P@10 = 0.2800, Rprec = 0.1527. And for the random baseline the

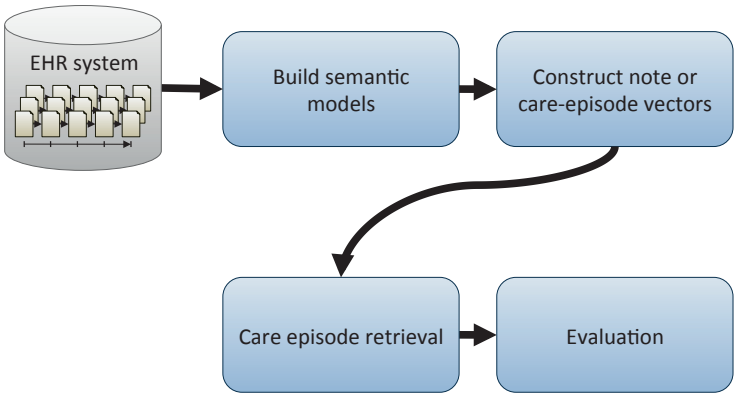

Figure 3.2: Illustration of the care episode retrieval experiments in Paper C. This is Figure 2 in Moen et al. (2015), Paper C in this thesis.

scores were: MAP = 0.0178, P@10 = 0.0175, Rprec = 0.0172.

3.3.2 Retrospective View and Results/Contributions

This paper was a continuation of the work presented in Moen et al. (2014b). This study focuses on exploring a set of distributional semantic models for use in retrieval of care episodes, only relying on the free-text information therein. One motivation for conducting this research was that we wanted to use this type of care episode retrieval in the summarization task presented in Paper E.

Such multi-document (multi-note) information retrieval — where also the query is a care episode, i.e., a collection of documents/notes — is a rather unique task as far as I know. Naturally, finding a reliable way of conducting automated evaluation was a central issue here. The number of result tables became fairly large due to the fact that we were evaluating eight different models/systems on two different evaluation setups. The results indicate that using ICD-10 codes as context for training the semantic models seems to be a promising direction for information retrieval on the level of care episodes. It is likely that using other training features from clinical practice, commonly documented in EHRs, could potentially produce even better semantic spaces for this task, and in general — models more suited for information access and NLP in the clinical domain. The use of such domain-specific context features in constructing semantic models would also be interesting to evaluate on the level of word synonyms, e.g., similar to the experiment in Paper A.

3.4 Paper D: On Evaluation of Automatically Generated Clinical Discharge Summaries

Authors: *Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski and Sanna Salanterä*

3.4.1 Summary

Proper evaluation is crucial for developing high-quality computerized text summarization systems. In the clinical domain, the specialized information needs of the clinicians complicate the task of evaluating automatically constructed text summaries — constructed from the free-text information that clinicians document in relation to patients' care episodes. The focus of this paper is on evaluation. We propose an automated and manual evaluation approach. We are not interested in the actual performance of some summarization method, instead the focus is on determining if, and to what degree, there is a correlation between how the automated and manual evaluation approaches rank various summarization methods. We assume that the manual evaluation scores are good indicators for relative performance, however this is not clear for the automated evaluation measures in question. Further, if such a correlation is observed, we may rely on the much faster automated evaluation when further developing the summarization methods. The experiment setup is illustrated in Figure 3.3.

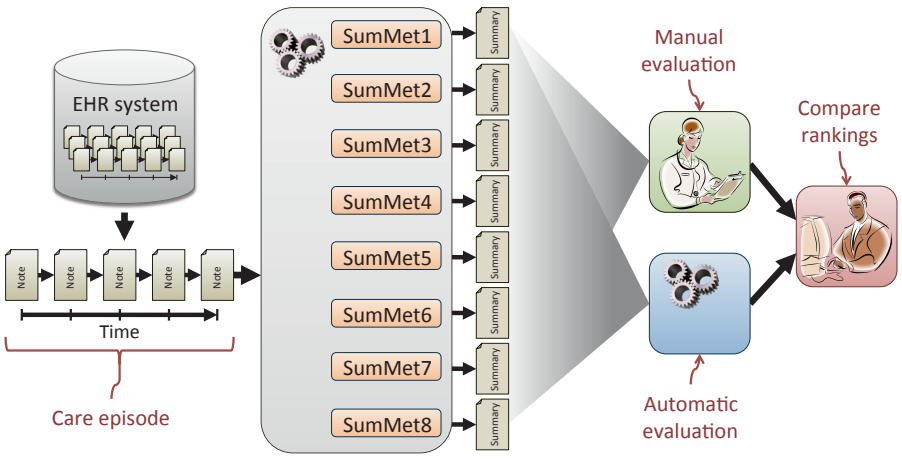

Figure 3.3: Illustration of the evaluation experiment conducted in Paper D.
This is a slightly modified version of Figure 1 in Moen et al. (2014a), Paper D in this thesis.

Four different evaluation measures in the *ROUGE evaluation toolkit* are explored in the automated evaluation approach, where the utilized gold standard for each summarized care episode is the accompanying discharge summary. The manual evaluation is performed by domain experts who use an evaluation scheme/tool that we developed as part of this study. The scores from the manual evaluation is calculated from the average summarization method scores from five care episodes. The scores from the automatic evaluation is based on the average scores from 156 care episodes. To identify which of the automatic evaluation metrics that best follows the manual evaluation, Pearson product-moment correlation coefficient (PPMCC) and Spearman's rank correlation coefficient (Spearman's rho) were calculated between the normalized manual evaluation scores and each of the automatic evaluation scores.

We find that all ROUGE measures correlate well with that of the manual evaluation, where the ROUGE-SU4 measure correlates the most. It achieves: PPMCC = 0.9510 (p-value = 0.00028), and Spearman's rho score = 0.8571 (p-value = 0.00653). The agreement among the manual evaluators is "good" according to guidelines on interrater agreement. These preliminary results indicate that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance. This allows us to rely on the presented automatic evaluation approach when further developing automatic text summarization for clinical text.

3.4.2 Retrospective View and Results/Contributions

A central issue in the works related to this thesis has been to enquire evaluation data suited for evaluating the different methods that have been introduced along the way. Automatic text summarization is in itself a very complex and challenging task, and to assess what is a good or poor summary is heavily influenced by the perspective of the judging subject and the underlying task. This is also the case when it comes to the task of generating and evaluating clinical free-text summaries. To *automate* such evaluation complicates things further. These are the reasons why in this paper we chose to try to generate a summary that is comparable to discharge summaries from care episodes — which clinicians construct/write manually as a part of the patient discharge process. This enables the use of hospital guidelines for manual judgement of the content and quality of a summary (c.f. the utilized manual evaluation scheme/tool). Further, this enables us to use the original discharge summaries as gold standard when performing automated evaluation. However, the evaluation approach does not provide any absolute truth when it comes to how the summarization methods performs, but primarily how they perform in relation to each other.

The main results of this experiment was that a correlation was found between: a) how human evaluators rank a set of different summarization methods, and b) how some automated evaluation metrics rank these same summarization methods. With this knowledge we could rely on the automated evaluation approach for rapid evaluation during further experimentation and development of summarization methods. This was a important step in the process of developing some of the summarization methods presented in Paper E.

3.5 Paper E: Comparison of automatic summarisation methods for clinical free text notes

Authors: *Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski and Sanna Salanterä*

3.5.1 Summary

Managing the information in EHR systems tends to be time consuming for clinicians (Farri et al. 2012, Hirsch et al. 2015). Automatic text summarization could assist in providing an overview of the free-text information in ongoing or finished care episodes, as well as in writing the final discharge summaries. This work focuses on summarization of the clinical free text written by clinicians (physician) in care episodes. We evaluated eight different automatic text summarization methods. Among these are four novel extraction-based text summarization methods, tailored for summarizing the free-text content of care episodes. A key feature of these methods is that they try to take into account the sequential and repetitive nature of the documented text/information. Most of them rely on the use of distributional semantic models, exploiting various textual features found in care episodes.

Care episodes used in this study are from EHRs belonging to heart patients admitted to a university hospital in Finland. The performance of the summarization methods are evaluated both automatically and manually. We utilized the *ROUGE evaluation toolkit* for automatic evaluation with discharge summaries used as gold standard, while a rating-based evaluation scheme/tool is used for the manual evaluation. By comparing how the automatic and manual evaluations correlates in terms of how they rank the different summarization method, we are able to perform a meta-evaluation of these ROUGE evaluation measures. Figure 3.4 shows an overview of the experimental setup.

The results show that there is a good agreement between the manual evaluators.

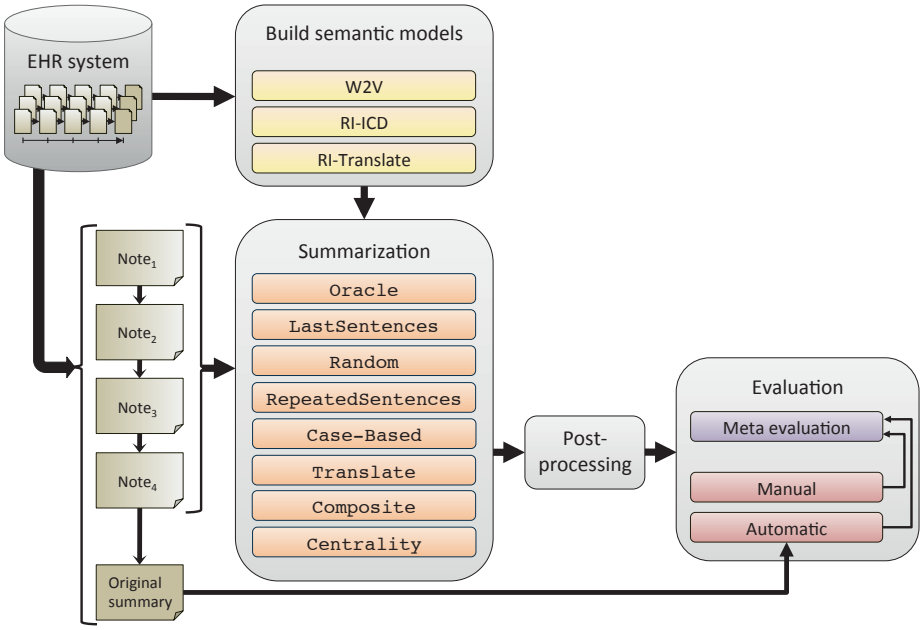

Figure 3.4: Illustration of the text summarization experiments conducted in Paper E. This is a slightly modified version of Figure 1 in Moen et al. (2016), Paper E in this thesis.

There is also a high correlation between how the manual evaluators and the automated evaluation rank the various summarization methods. Here the ROUGE-N2 and ROUGE-l metrics have the highest correlation with the manual evaluators. The high correlation between manual and automated evaluations suggests that the less labor-intensive automated evaluations can be used as a proxy for human evaluations when developing summarization methods. This is of significant practical value for summarization method development aimed at this task. Both the automated and manual evaluations agree in that a proposed composition based summarization method outperforms all the other considered methods.

3.5.2 Retrospective View and Results/Contributions

Here the focus was on exploring a set of different methods, exploring various features of clinical free-text information, for performing the automated summarization. They all rely solely on exploiting statistical features that are found in EHRs/care episodes, without the use of any manual annotation or similar manual work tailored for this task. This again reflects the underlying motivation for our approach; to explore ways of conducting such summarization while surpassing the need for developing NLP resources tailored for the task.

The utilized automatic evaluation approach was the same as that used in Paper D. However, for the manual evaluation, a somewhat simplified evaluation scheme (compared to that used in Paper D) was used by the human evaluators. The evaluators found the original evaluation scheme from Paper D to be very time-consuming to use due to its complex , thus a simplification was done in order to allow for evaluation of more summaries within certain time and resource limits. Yet, the results showed that a correlation between the manual and automatic evaluation was present also in this experiment.

With today's hospital practice, the optimal text summary generated from care episodes through sentence-level extraction-based text summarization will hardly ever become identical to a corresponding discharge summary written by a clinician. One reason for this is that much of the information that clinicians write in a discharge summary is never written in the clinical (daily) notes that constitute a care episode. However, we demonstrate here that there are certain sentence-level textual features that can be indicative of sentences inclusion potential in a (discharge) summary.

Future work on this task includes further developing these methods so that they can be used for assisting clinicians through semi-automatic, user-guided, discharge summary writing. There are of course also other summarization approaches and methods that should be explored, including exploiting other (statistical) features of care episodes and clinical text.

The summarization methods explored in this paper are potentially suited for presenting an *indicative overview* of the free-text content written in a — possible ongoing — care episode, that clinicians could read in situations where they do not have time to read through all previously written clinical notes. This would be supplementary to a comprehensive overview/visualization of the more structured and coded data in EHRs, such as images, laboratory test results, medications, diagnosis codes (see Roque et al. (2010), Pivovarov and Elhadad (2015)). Further, situations where they could find use would be where the need for such an overview outweighs the possible patient safety issues that may be caused by lack of relevant information in the generated summary.

Future work should focus more towards conducting *extrinsic evaluation* — evaluating how the use of automatic free-text summarization systems in a (simulated) clinical setting will impact documentation speed and quality, as well as health care quality and patient outcomes.

Chapter 4

Conclusions and Recommendations for Future Work

This thesis has focused on distributional semantic methods used to construct domain and task specific semantic models from primarily clinical text. Five sets of experiments have been presented, published as separate papers, that focused on different applications of distributional semantic models. Three of these focus on textual similarity assessment on different granularity levels: words (Paper A), sentences (Paper B), and clinical notes (Paper C), where some novel ways of training the utilized distributional semantic models were presented and evaluated. The other two papers, Papers D and E, focus on applications of semantic models in free-text summarization methods tailored for clinical text, as well as evaluations of these. Both existing and novel approaches and methods have been applied and evaluated in these experiments.

4.1 Conclusions

Three research questions were presented in Section 1.2. Here these are linked to the experiments in the various papers, discussed and concluded.

Research Question 1

How can the distributional hypothesis be utilized in constructing semantic similarity models suited for clinical text?

In Paper A we found that combining distributional semantic models trained dif-

ferently can yield improved performance in terms of modeling word synonym relations. First, combining the retrieved candidate words from two models (Random Indexing (RI) and Random Permutation (RP)), both trained on the same corpus seemingly enhances the synonymic relations between the query and the resulting top extracted candidate words, outperforming the use of one model alone. Second, combining four models, trained using RI and RP on one of two different training corpora — one consisting of clinical text and the other medical research literature — improves the results further ($\text{RI}_{clinical}$ + $\text{RI}_{medical}$ + $\text{RP}_{clinical}$ + $\text{RP}_{medical}$). This also outperforms the approach of using a conjoint corpus for training. This suggests that such a multi-model approach allows for a broader range of semantic similarity features to be captured from free-text corpora, and that combining the models in various ways (their cosine similarity scores) may elevate certain desirable semantic similarity features.

In Paper B the use of multiple sense-vectors per word is explored as a possible approach to enhancing the spectrum of semantic relations captured in the resulting semantic model. Although the presented method demonstrated only minor improvements over the classical RI training approach, the work on multi-sense semantic similarity methods and models calls for further research. One possible use is in sentence similarity assessment and sentence topic clustering for use in extraction-based text summarization, similar to how it is done in Paper E. This could for example assist in providing more fine-grained discrimination between which sentence-topic cluster each sentence should belong to. Also, relatively little work is published on the use of such word-level multi-sense semantic similarity models in compositionality for applications including composing sentence context vectors and document context vectors.

Training of the distributional semantic models used in Papers A and B is done using a sliding window approach, where the context is defined by the neighboring words in the training corpora. However, in Paper C we explore also the use of other contextual features for training. Most notable are the methods relying on labeled ICD-10 codes and their internal hierarchy as context for inducing word-level semantics. This achieves the best results compared to the other evaluated methods and systems. As the experiment focused on care episode similarity (care episode retrieval), it is arguably natural that the use of such domain-specific meta-information as training features results in semantic spaces that better reflect semantic similarity relations suited for this task. However, the use of domain-specific (meta) features for constructing semantic models is something that deserves to be explored further, possibly evaluated on other granularity levels, such as word and sentence similarity assessment. For instance, this may include medications, allergies, age, lab tests, relevant clinical practice guidelines, SNOMED-CT concepts

and so on, alone or in combination with neighboring words (c.f. sliding window). Paper C also includes the use of word2vec (W2V) as an alternative to RI for constructing distributional semantic models. The results show that W2V-based models outperform RI-based ones when constructed/trained using comparable context features and identical corpora.

Research Question 2

What sentence-level features of clinical text in care episodes are indicative of relevancy for inclusion in a clinical free-text summary?

Based on the experiments in Paper E we can conclude the following:

- Clustering sentences into topics that span across clinical notes is a seemingly desirable way of reducing redundancy with respect to what is of interest to the reader (clinician).
- The importance of a sentence in a clinical note is related to how many times the same/similar information has been mentioned throughout a care episode.
- By looking at discharge summaries from other similar care episodes, one can assess the importance of a sentence based on whether or not the same or similar information has been written there.
- If, using a VSM-based translation system, a sentence (its vector representation) can be "translated" into a vector representation that is similar to how this same sentence would look like in the translated vector space, it should be considered for inclusion in the final summary.

The experiments in the thesis represents some initial steps towards the goal of enabling summarization of care episodes in an fully unsupervised and resource light manner. It is not evident just how far one can go with the selected approach. A more user-centered evaluation is needed in order to shed additional light on the strengths, weaknesses and limitations of the explored summarization methods.

Research Question 3

How can the evaluation of distributional semantic models and text summaries generated from clinical text be done in a way that is fast, reliable and inexpensive?

Evaluation setups that allow for rapid automated evaluation are crucial when developing new computerized methods and algorithms. The most common way to do this is to manually construct gold standards, such as those made in relation to

various shared tasks. In the synonym extraction task in Paper A, we used a set of MeSH terms and their synonyms as a gold standard. However, when conducting a manual analysis of extracted samples, we found that the semantic models not only extract synonyms that are present in the gold standard, but also other equally valid synonyms not present there. This indicates that constructing a complete list of synonyms for a word/term is challenging, especially when its usage is not clearly defined in terms of context and (sub-)domain. Further, this indicates that distributional semantic models can be used to improve coverage of lexical resources.

In Paper C we conclude that experiments conducted in most of the related work, including ours, are based on evaluation through pure retrieval performance according to a predefined gold standard. This is normally referred to as *intrinsic evaluation* (Hirschberg and Manning 2015). Future research on information retrieval in the clinical domain should arguably focus more on *extrinsic evaluation* — evaluating information retrieval systems in terms of support for health care and patient outcomes, as also argued in Mishra et al. (2014), Hirschberg and Manning (2015).

The evaluations conducted in Papers D and E are also defined as *intrinsic* in that pre-defined gold standards are used as evaluation criteria. Here we suggest that future work on such a task should (also) incorporate extrinsic evaluation — evaluating how the use of automatic text summarization systems in a clinical setting will impact documentation speed and quality, as well as health care quality and patient outcomes.

4.2 Future Work

Through the various experiments presented in this thesis, a number of possibilities for future work have been unveiled. The following are suggestions for future work in the context of the three research questions and related experiments

A major focus has been the training and use of distributed semantic models, where several novel methods for constructing such models with various properties have been proposed and evaluated. The main use of these models has been to compute semantic textual similarity between linguistic items of various granularity (words, sentences, clinical notes/care episodes).

For training there are multiple interdependent parameters and training features involved, these are mainly: the dimensionality of the VSM (fixed or post-training reduced); what training approach to use for inducing the vector space (RI vs RP vs W2V CBOW vs W2V Skip-gram); what contextual features to use and how to weight these (e.g., using sliding window with a certain window size); non-zero elements for the index vectors used in the RI approach; thresholds concerning

clustering, such as the thresholds used in the sentence-topic clustering for text summarization and in the sense clustering in the multi-sense RI method; lower and upper frequency filters, e.g., filtering out words that occur less or more than some given thresholds. An additional factor is the utilized training corpus: the type of text and domain, or domains; the(ir) size; how pre-processing should be done (tokenization, lemmatization/stemming, stop-word removal, etc.). There is little doubt that improvements can be gained through optimizing these parameters and training features.

However, it is not evident how such optimization should be done since there is virtually an infinite number of different parameter values, possibly training features and combinations that can be tested and explored. Nor is it completely evident how different parameter settings and training features are related or how they may affect the resulting models and task performance. This is linked to the unsupervised nature of the underlying data-driven training approach together with the vast complexity of the training data (natural language text) and its size (millions of documents). Further, it is likely that there are no universally optimal settings, but optimal settings are instead task-specific. Thus there is a relatively long "distance" between setting the initial parameter values to having a fully trained distributional semantic model that has been properly evaluated on a given task. This is particularly the case when the task has a certain complexity level to it (e.g. text summarization). Henriksson and Hassel (2013) explore various vector dimensionalities using a fixed set of value alternatives. A somewhat similar type of exploration was conducted in Moen and Marsi (2013). Future work in this direction could focus more on evaluating the different parameter values as a multi-variable optimization task, possibly using some type of *gradient*-based or *hill-climbing* search algorithm (Russell and Norvig (2005), page 139 and 149). A possible outcome may include suggestions for how such parameter optimization should be approached.

For languages where comprehensive lexical knowledge resources are available, such as the SNOMED-CT and WordNet ontologies for English, hybrid approaches that combine statistical semantics with lexical resources, e.g., using the approach by Faruqui et al. (2015), could potentially contribute to producing semantic spaces that more correctly reflects the clinical terminology. Also for tasks where proper evaluation data is available, some type of task-specific retrospective fitting of pre-constructed distributional semantic models could be performed. For example, one may explore a similar approach as that used by Chen and Manning (2014), where they explore the use of a pre-trained neural network language model as the starting point for training a neural network classifier for use in a dependency parser. This general direction implies "moving" some of the context vectors in a semantic model so that they better reflect the (semantic) similarity relations expected by the

ontologies or training examples. Ideally this would also result in movement of the context vectors belonging to neighboring words and concepts that are not found in the ontologies or training examples.

Semantic vectors representing sentences, notes or care episodes have here primarily been composed from word context vectors. This was done essentially through pointwise summation of the constituent (normalized) word context vectors. One obvious drawback with this approach is that word order is not taken into consideration. For instance, given a sentence, it is obvious that word order is of significance to the meaning it is meant to convey (e.g., "dog eats rabbit"). However, Landauer et al. (1997) concludes that when grading similarity between texts (essays), word order is seemingly not of great importance when relying on a distributional semantic model (LSA) to compute similarities. However, given the task of computerized semantic textual similarity assessment, where our strategy is to compose sentence context vectors for judging semantic similarity between sentence pairs. Here we would expect to see that improvements over classic distributional semantic models will be achieved by models and composition techniques that to some degree are able to produce sentence vectors whose point in the semantic space is adjusted based on the order of its constituent words (e.g., $\overrightarrow{dog} + \overrightarrow{eats} + \overrightarrow{rabbit}$ VS $\overrightarrow{rabbit} + \overrightarrow{eats} + \overrightarrow{dog}$). This would differ from approaches where training is done based on, e.g., co-occurrence information of pre-defined phrases, or where some type of convolution or shifting is used to construct completely new vectors. Perhaps a plausible approach would share certain similarities with multi-sense modeling techniques.

In looking at the work by Tversky (1977), one may argue that the use of distributional semantics for semantic textual similarity assessment is somewhat comparable to how humans judge similarity between concepts or objects. That is, similarity between two concepts is calculated based on measuring the likeness of, and the lack of, common features. However, Tversky also shows that the context in which the objects are evaluated has an impact on human similarity assessment. Further, the order in which two objects are compared may have an impact on how similarity is judged, i.e., $sim(x, y) \neq sim(y, x)$. The latter two properties of similarity are today not well explored in work on distributional semantics. I am not aware of any published work on bi-directional VSMs or distance metrics. Enabling this, as well as general improvements to automatically induced semantics, it may require that new ways of training and representing semantic models have to be introduced. This includes representations that can model (domain-/corpus-specific) bi-directionality and training algorithms that captures these properties from distributional statistics in text. Such a representation may, for example, be in the form of a multi-dimensional (hyper-)cube, or some type of graph. Regarding capturing

these properties, it may be so that we have to identify and explore alternatives to the *distributional hypothesis*, alternatives that are based on some other fundamental properties of language and text, yet applicable to fully unsupervised methods for learning semantics.

In the present work on automatic text summarization, a set of features were explored in terms of their (statistical) indication of sentences relevance in clinical free-text summaries. The explored features are primarily based on statistics about (sequential) information redundancy and what other clinicians have deemed important in comparable cases. An exploratory approach was used when investigating potential features, motivated primarily by: more and less obvious patterns observable in clinical notes; observations reported in related work; feedback and suggestions from domain experts. Better understanding of such features could potentially be gained through conducting a thorough observation of domain experts during their work, in particular the actual process of writing or constructing summarized information and discharge summaries.

It is still unclear how far one can go with such an unsupervised approach, compared to some type of knowledge modeling. The latter could, e.g., include manual identification and classification of the significance of pre-defined concepts and features in clinical text, and the extraction of these to construct a summary, similarly to how it is done by Velupillai and Kvist (2012). However, this introduces the need for extensive manual labor. Hybrid approaches — that combines significance scoring of concepts derived from both supervised and unsupervised methods — is something to explore in future work.

Another important part of information summarization is how the summarized information is presented to the user. This has not been a focus in the work in this thesis, but will be a natural part of future work.

The evaluation criteria used in the automatic text summarization work was mainly based on discharge summaries. For automated evaluation, the gold standard consisted of the original discharge summaries, while for manual evaluation we used hospital guidelines concerning discharge summary content. This evaluation gave some indications of how well we are able to "reproduce" the content of a discharge summary. However, both Mishra et al. (2014) and Hirschberg and Manning (2015) conclude that future research should focus more on evaluating the impact of introducing text summarization systems in (simulated) hospital settings. I believe that this is the natural next step with respect to evaluation of this text summarization work. Such an evaluation, with clinicians as users of the system, is likely to provide more qualitative feedback regarding performance and user needs. This could also provide indications for features to use in terms of assessing information sig-

nificance relative to the summarization process.

Also, I believe that a more user-guided summarization system is required, in particular when it comes to supporting the process of writing discharge summaries. One approach could include having the summarization system iteratively suggest sentences for inclusion based on first analyzing what content the user has, at any point, written in, or extracted into, the summary under creation.

When looking at the evaluation conducted in the other presented experiments (Papers A, B and C), the used gold standards consist of classifications done by humans. However, in Paper A it was revealed that the gold standard used for synonyms was lacking in terms of coverage since it lacked true positives. It is likely that this is also the case for the evaluation setup used in Paper C. This shows that a complete and clear cut gold standard is challenging to produce when dealing with natural language text. However, there are few alternative evaluation approaches available that also support rapid evaluation during method development. Chapman (2010) argue that it is important to involve end-users early (earlier) in the development process of NLP applications designed for assisting in patient care. Likewise, I believe that future work on development and evaluation of (distributional) semantic models for use in clinical NLP applications, would benefit from incorporating the end-users at various stages in the process. This could assist the developers greatly when it comes to understanding the domain and what (con)textual features that could potentially be utilized to achieve the desired semantic relations and properties in the resulting model.

4.3 Final Remarks

The work presented in this thesis could be of inspiration to others when it comes to constructing distributional semantic similarity models for use in the clinical domain. Several methods have been presented and evaluated at various textual semantic similarity assessment tasks, primarily using clinical text. We have also seen an approach to automated summarization of clinical free-text information that primarily relies on the (re-)use of statistical information derived from clinical corpora. This direction, focusing on the reuse of the large amounts of digitally stored clinical data being accumulated in hospitals nowadays, could facilitate the development of resource-lean software tools able to support and ease the information access and management work for clinicians and others in the health sector.

References

Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, volume 3491 of *Lecture Notes in Computer Science*. Springer-Verlag, Bath, England, 2004.

Aasen, Sigrun Espelien. News from the MeSH special interest group; MeSH speaks Norwegian in 2013! *Journal of the European Association for Health Information and Libraries*, 9(1):38–40, 2012.

Abulkhair, Maysoon; ALHarbi, Nora; Fahad, Amani; Omair, Seham; ALHosaini, Hadeel, and AlAffari, Fatimah. Intelligent integration of discharge summary: A formative model. In *Intelligent Systems Modelling & Simulation (ISMS 2013), 4th International Conference on Intelligent Systems, Modelling and Simulation*, pages 99–104, Bangkok, Thailand, 2013. Institute of Electrical and Electronics Engineers.

Afantenos, Stergos; Karkaletsis, Vangelis, and Stamatopoulos, Panagiotis. Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

Agirre, Eneko; Cer, Daniel; Diab, Mona, and Gonzalez-Agirre, Aitor. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 385–393, Montreal, Canada, June 2012. Association for Computational Linguistics.

Agirre, Eneko; Cer, Daniel; Diab, Mona; Gonzalez-Agirre, Aitor, and Guo, Weiwei. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint*

*Conference on Lexical and Computational Semantics (*SEM)*, volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Allvin, Helen; Carlsson, Elin; Dalianis, Hercules; Danielsson-Ojala, Riitta; Daudaravičius, Vidas; Hassel, Martin; Kokkinakis, Dimitrios; Lundgren-Laine, Heljä; Nilsson, Gunnar; Nytrø, Øystein; Salanterä, Sanna; Skeppstedt, Maria; Suominen, Hanna, and Velupillai, Sumithra. Characteristics and analysis of Finnish and Swedish clinical intensive care nursing narratives. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 53–60, Los Angeles, California, USA, 2010. Association for Computational Linguistics.

Ao, Hiroko and Takagi, Toshihisa. ALICE: An Algorithm to Extract Abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 12 (5):576–586, 2005.

Appelt, Douglas E and Onyshkevych, Boyan. The common pattern specification language. In *Proceedings of the TIPSTER workshop*, pages 23–30, Baltimore, Maryland, USA, 1998. Association for Computational Linguistics.

Ashburner, Michael; Ball, Catherine A; Blake, Judith A; Botstein, David; Butler, Heather; Cherry, J Michael; Davis, Allan P; Dolinski, Kara; Dwight, Selina S; Eppig, Janan T; Harris, Midori A; Hill, David P; Issel-Tarver, Laurie; Kasarskis, Andrew; Lewis, Suzanna; Matese, John C; Richardson, Joel E; Ringwald, Martin; Rubin, Gerald M, and Sherlock, Gavin. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Baroni, Marco and Lenci, Alessandro. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Baroni, Marco; Dinu, Georgiana, and Kruszewski, Germán. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247. Association for Computational Linguistics, 2014.

Bartlett, Christopher; Boehncke, Klaus, and Haikerwal, M. E-health: Enabler for Australia's health reform. *Melbourne: Booz & Co*, 2008.

Bashyam, Vijayaraghavan; Hsu, William; Watt, Emily; Bui, Alex A. T.; Kangarloo, Hooshang, and Taira, Ricky K. Problem-centric organization and visualization of patient imaging and clinical data. *RadioGraphics*, 29(2):331–343, 2009.

Bateman, John A; Magnini, Bernardo, and Fabris, Giovanni. The generalized upper model knowledge base: Organization and use. In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, pages 60–72. IOS Press, Amsterdam, 1995.

Berman, Jules J. Doublet method for very fast autocoding. *BMC Medical Informatics and Decision Making*, 4(1):16, 2004.

Blei, David M.; Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Blumenthal, David and Tavenner, Marilyn. The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504, 2010.

Bullinaria, John A and Levy, Joseph P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, 44(3):890–907, 2012.

Carbonell, Jaime and Goldstein, Jade. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne, Australia, 1998. Association for Computing Machinery.

Chapman, WW. Closing the gap between NLP research and clinical practice. *Methods of Information in Medicine*, 49(4):317, 2010.

Chatterjee, Niladri and Mohan, Shiwali. Extraction-based single-document summarization using Random Indexing. In *19th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2007*, volume 2, pages 448–455, Patras, Greece, 2007. Institute of Electrical and Electronics Engineers.

Chen, Danqi and Manning, Christopher. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics.

Chute, Christopher G. Classification and retrieval of patient records using natural language: An experimental application of Latent Semantic Analysis. In *Engineering in Medicine and Biology Society, Proceedings of the Annual International Conference of the IEEE*, volume 13, pages 1162–1163, Orlando, Florida, USA, 1991. Institute of Electrical and Electronics Engineers.

Cohen, Paul R and Howe, Adele E. How evaluation guides AI research: The message still counts more than the medium. *AI magazine*, 9(4):35, 1988.

Cohen, Paul R and Howe, Adele E. Toward AI research methodology: Three case studies in evaluation. *Systems, Man and Cybernetics, IEEE Transactions on*, 19 (3):634–646, 1989.

Cohen, Raphael; Aviram, Iddo; Elhadad, Michael, and Elhadad, Noémie. Redundancy-aware topic modeling for patient record notes. *PloS one*, 9(2): e87555, 2014.

Cohen, Trevor. Exploring MEDLINE space with random indexing and pathfinder networks. In *AMIA Annual Symposium Proceedings*, volume 2008, pages 126–130, Washington, DC, USA, 2008. American Medical Informatics Association.

Cohen, Trevor and Widdows, Dominic. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2): 390–405, April 2009.

Cohen, Trevor; Schvaneveldt, Roger, and Widdows, Dominic. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240–256, 2010.

Collobert, Ronan; Weston, Jason; Bottou, Léon; Karlen, Michael; Kavukcuoglu, Koray, and Kuksa, Pavel. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Cutting, Doug. Apache Lucene open source package. `http://lucene.apache.org/`, 1999. (accessed 1st March 2016).

Dalianis, Hercules; Hassel, Martin, and Velupillai, Sumithra. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*, pages 243–249, 2009.

De Vine, Lance; Zuccon, Guido; Koopman, Bevan; Sitbon, Laurianne, and Bruza, Peter. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information*

and Knowledge Management, pages 1819–1822, Shanghai, China, 2014. Association for Computing Machinery.

Demner-Fushman, Dina; Chapman, Wendy W, and McDonald, Clement J. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.

Erkan, Günes and Radev, Dragomir R. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

European Commission,. eHealth Action Plan 2012-2020:
Innovative healthcare for the 21st century.
https://ec.europa.eu/digital-single-market/en/news/ehealth-action-plan-2012-2020-innovative-healthcare-21st-century, 2012. (accessed 1st March 2016).

Farri, Oladimeji; Pieckiewicz, David S; Rahman, Ahmed S; Adam, Terrence J; Pakhomov, Serguei V, and Melton, Genevieve B. A qualitative analysis of EHR clinical document synthesis by clinicians. In *AMIA Annual Symposium Proceedings*, pages 1211–1220, Chicago, IL, USA, 2012. American Medical Informatics Association.

Faruqui, Manaal; Dodge, Jesse; Jauhar, Sujay Kumar; Dyer, Chris; Hovy, Eduard, and Smith, Noah A. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Feng, Jin; Zhou, Yi-Ming, and Martin, Trevor. Sentence similarity based on relevance. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 8, pages 832–839, Málaga, Spain, 2008.

Ferreira, Rafael; Lins, Rafael Dueire; Simske, Steven J.; Freitas, Fred, and Riss, Marcelo. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, 39:1–28, 2016. ISSN 0885-2308.

Ferrucci, David; Brown, Eric; Chu-Carroll, Jennifer; Fan, James; Gondek, David; Kalyanpur, Aditya A; Lally, Adam; Murdock, J William; Nyberg, Eric; Prager, John; Schlaefer, Nico, and Welty, Chris. Building Watson: An overview of the DeepQA project. *AI magazine*, 31(3):59–79, 2010.

Frege, Gottlob. Über Sinn Und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50, 1892.

Friedman, Carol; Kra, Pauline, and Rzhetsky, Andrey. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, 2002.

Friedman, Carol; Rindflesch, Thomas C, and Corn, Milton. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of Biomedical Informatics*, 46(5):765–773, 2013.

Furnas, George W; Landauer, Thomas K; Gomez, Louis M, and Dumais, Susan T. Human factors and behavioral science: Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806, 1983.

Goldstein, Jade; Mittal, Vibhu; Carbonell, Jaime, and Kantrowitz, Mark. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, volume 4, pages 40–48. Association for Computational Linguistics, 2000.

Grabar, Natalia; Varoutas, Paul-Christophe; Rizand, Philippe; Livartowski, Alain, and Hamon, Thierry. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, 48(2):149, 2009.

Guevara, Emiliano. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 135–144. Association for Computational Linguistics, 2011.

Hahn, Udo and Mani, Inderjeet. The challenges of automatic summarization. *Institute of Electrical and Electronics Engineers - Computer*, 33(11):29–36, 2000.

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

Harris, Zellig S. Distributional structure. *Word*, 10:146–162, 1954.

Hassel, Martin and Sjöbergh, Jonas. Navigating through summary space: Selecting summaries, not sentences. In *Resource Lean and Portable Automatic Text Summarization*, pages 109–132. KTH Royal Institute of Technology, 2007.

Henriksson, Aron and Hassel, Martin. Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In *Proceedings of the Louhi Workshop on Health Document Text Mining and Information Analysis*, pages 1–6, 2013.

Henriksson, Aron; Conway, Mike; Duneld, Martin, and Chapman, Wendy Webber. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *AMIA 2013, American Medical Informatics Association Annual Symposium*, Washington, DC, USA, 2013a.

Henriksson, Aron; Skeppstedt, Maria; Kvist, Maria; Duneld, Martin, and Conway, Mike. Corpus-driven terminology development: populating Swedish SNOMED-CT with synonyms extracted from electronic health records. In *Proceedings of BioNLP*, pages 36–44, Sofia, Bulgaria, 2013b. Association for Computational Linguistics.

Hevner, Alan R; March, Salvatore T; Park, Jinsoo, and Ram, Sudha. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

Hirsch, Jamie S; Tanenbaum, Jessica S; Lipsky Gorman, Sharon; Liu, Connie; Schmitz, Eric; Hashorva, Dritan; Ervits, Artem; Vawdrey, David; Sturm, Marc, and Elhadad, Noémie. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2015.

Hirschberg, Julia and Manning, Christopher D. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. Association for Computing Machinery.

Huang, Eric H.; Socher, Richard; Manning, Christopher D., and Ng, Andrew Y. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Jha, Ashish K. Meaningful use of electronic health records: The road ahead. *The Journal of the American Medical Association*, 304(15):1709–1710, 2010.

Johnson, William B and Lindenstrauss, Joram. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Jones, Karen Sparck. Automatic summarising: Factors and directions. *Advances in Automatic Text Summarization*, pages 1–12, 1999.

Jonnalagadda, Siddhartha; Cohen, Trevor; Wu, Stephen, and Gonzalez, Graciela. Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129–140, 2012.

Kanerva, Pentti; Kristofersson, Jan, and Holst, Anders. Random Indexing of text samples for Latent Semantic Analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036, Philadelphia, PA, USA, 2000.

Karlgren, Jussi and Sahlgren, Magnus. From words to understanding. In *Foundations of Real World Intelligence*, CSLI Lecture Notes 125, pages 294–308. CSLI, 2001.

Kate, Rohit J. Unsupervised grammar induction of clinical report sublanguage. *Journal of Biomedical Semantics*, 3(S-3):S4, 2012.

Kolb, Peter. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference on Computational Linguistics, NODALIDA'09*, volume 4, pages 81–88, Odense, Denmark, 2009. NEALT Proceedings series.

Koopman, Bevan; Zuccon, Guido; Bruza, Peter; Sitbon, Laurianne, and Lawley, Michael. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2439–2442, Maui, HI, USA, 2012. Association for Computing Machinery.

Kripalani, Sunil; LeFevre, Frank; Phillips, Christopher O; Williams, Mark V; Basaviah, Preetha, and Baker, David W. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *Journal of the American Medical Association*, 297(8):831–841, 2007.

Kuhn, Harold W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Kvist, Maria; Skeppstedt, Maria; Velupillai, Sumithra, and Dalianis, Hercules. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems-future vision, a physician's perspective. In *9th Scandinavian Conference on Health Informatics, SHI 2011*, Oslo, Norway, 2011. Tapir Academic Press.

Landauer, Thomas K and Dumais, Susan T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211, 1997.

Landauer, Thomas K; Laham, Darrell; Rehder, Bob, and Schreiner, Missy E. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417. Cognitive Science Society, Citeseer, 1997.

Le, Quoc and Mikolov, Tomas. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, Beijing, China, 2014.

LeCun, Yann; Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Lehman, Ann. *JMP for basic univariate and multivariate statistics: A step-by-step guide*. SAS Institute, 2005.

Lenci, Alessandro and Benotto, Giulia. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 75–79, Montreal, Canada, 2012. Association for Computational Linguistics.

Levy, Omer and Goldberg, Yoav. Linguistic regularities in sparse and explicit word representations. In *Proceedings of Eighteenth Conference on Computational Natural Language Learning (CoNNL-2014)*, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

Levy, Omer; Goldberg, Yoav, and Dagan, Ido. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

Lin, Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Lissauer, T; Paterson, CM; Simons, A, and Beard, RW. Evaluation of computer generated neonatal discharge summaries. *Archives of Disease in Childhood*, 66 (4 Spec No):433–436, 1991.

Liu, Shuhua. Experiences and reflections on text summarization tools. *International Journal of Computational Intelligence Systems*, 2(3):202–218, 2009.

Lord, Phillip W.; Stevens, Robert D.; Brass, Andy, and Goble, Carole A. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

Luhn, Hans Peter. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

Lund, Kevin and Burgess, Curt. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, June 1996.

Manning, Christopher D; Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*, volume 1. Cambridge University Press, Cambridge, UK, 2008.

Marsi, Erwin; Moen, Hans; Bungum, Lars; Sizov, Gleb; Gambäck, Björn, and Lynum, André. NTNU-CORE: Combining strong features for semantic similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 66–73, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

McClelland, James L; Rumelhart, David E, and Group, PDP Research. Parallel distributed processing. *Explorations in the microstructure of cognition*, 2, 1986.

McCray, Alexa T; Srinivasan, Suresh, and Browne, Allen C. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 235–239, Washington, DC, USA, 1994. American Medical Informatics Association.

Mendonça, Eneida A; Haas, Janet; Shagina, Lyudmila; Larson, Elaine, and Friedman, Carol. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*, 38 (4):314–321, 2005.

Meng, Frank; Taira, Ricky K; Bui, Alex AT; Kangarloo, Hooshang, and Churchill, Bernard M. Automatic generation of repeated patient information for tailoring clinical notes. *International Journal of Medical Informatics*, 74(7):663–673, 2005.

Meystre, Stéphane M; Savova, Guergana K; Kipper-Schuler, Karin C, and Hurdle, John F. Extracting information from textual documents in the electronic health

record: A review of recent research. *Yearbook of Medical Informatics*, 35:128–44, 2008.

Mihalcea, Rada and Tarau, Paul. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 404–411, Barcelona, Spain, 2004. Association for Computational Linguistics.

Mikolov, Tomas; Chen, Kai; Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S., and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. 2013b.

Mikolov, Tomas; Yih, Wen-tau, and Zweig, Geoffrey. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, USA, June 2013c. Association for Computational Linguistics.

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

Mishra, Rashmi; Bian, Jiantao; Fiszman, Marcelo; Weir, Charlene R; Jonnalagadda, Siddhartha; Mostafa, Javed, and Del Fiol, Guilherme. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 2014.

Mnih, Andriy and Hinton, Geoffrey E. A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems (NIPS 2009)*, pages 1081–1088, Vancouver, B.C., Canada, 2009. Citeseer.

Moen, Hans and Marsi, Erwin. Cross-lingual random indexing for information retrieval. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 164–175. Springer Berlin Heidelberg, 2013.

Moen, Hans; Heimonen, Juho; Murtola, Laura-Maria; Airola, Antti; Pahikkala, Tapio; Terävä, Virpi; Danielsson-Ojala, Riitta; Salakoski, Tapio, and Salanterä, Sanna. On evaluation of automatically generated clinical discharge summaries. In *Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014)*, pages 101–114, Trondheim, Norway, 2014a. CEUR Workshop Proceedings.

Moen, Hans; Marsi, Erwin; Ginter, Filip; Murtola, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 116–124, Gothenburg, Sweden, 2014b. Association for Computational Linguistics.

Moen, Hans; Ginter, Filip; Marsi, Erwin; Peltonen, Laura-Maria; Salakoski, Tapio, and Salanterä, Sanna. Care episode retrieval: distributional semantic models for information retrieval in the clinical domain. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S2, 2015.

Moen, Hans; Peltonen, Laura-Maria; Heimonen, Juho; Airola, Antti; Pahikkala, Tapio; Salakoski, Tapio, and Salanterä, Sanna. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

Montague, Richard and Thomason, Richmond H. *Formal Philosophy: Selected Papers of Richard Montague; Ed. and with an Introduction by Richmond H. Thomason*. Yale University Press, 1976.

Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Citeseer, 2005.

Needleman, Saul B and Wunsch, Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

Neelakantan, Arvind; Shankar, Jeevan; Passos, Alexandre, and McCallum, Andrew. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

Nenkova, Ani and McKeown, Kathleen. *Automatic Summarization*. Foundations and Trends in Information Retrieval. Now Publishers Inc., 2011.

NLM, National Library of Medicine. MeSH (Medical Subject Headings), a. URL `http://www.ncbi.nlm.nih.gov/mesh`. (accessed 10th October 2015).

NLM, National Library of Medicine. International Health Terminology Standards Development Organisation: Supporting Different Languages, b. URL `http://www.ihtsdo.org/snomed-ct`. (accessed 10th October 2015).

NLM, National Library of Medicine. Unified Medical Language System, c. URL https://www.nlm.nih.gov/research/umls/. (accessed 10th October 2015).

Panman, Otto. Homonymy and polysemy. *Lingua*, 58(1):105–136, 1982.

Partee, Barbara; ter Meulen, Alice, and Wall, Robert. *Mathematical Methods in Linguistics*, volume 30. Springer Science & Business Media, 1990.

Patil, Kaustubh and Brazdil, Pavel. Sumgraph: text summarization using centrality in the pathfinder network. *International Journal on Computer Science and Information Systems*, 2(1):18–32, 2007.

Pedersen, Ted; Pakhomov, Serguei VS; Patwardhan, Siddharth, and Chute, Christopher G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

Pivovarov, Rimma and Elhadad, Noémie. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

Plate, Tony. Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI 1991)*, pages 30–35, San Mateo, CA, 1991. Citeseer.

Pradhan, Sameer; Elhadad, Noémie; Chapman, Wendy; Manandhar, Suresh, and Savova, Guergana. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University.

Pyysalo, Sampo; Ginter, Filip; Heimonen, Juho; Björne, Jari; Boberg, Jorma; Järvinen, Jouni, and Salakoski, Tapio. BioInfer: A corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50, 2007.

Pyysalo, Sampo; Ginter, Filip; Moen, Hans; Salakoski, Tapio, and Ananiadou, Sophia. Distributional semantics resources for biomedical text processing. In *Proceedings of Languages in Biology and Medicine*, 2013.

Rector, Alan L. Clinical terminology: why is it so hard? *Methods of Information in Medicine*, 38(4/5):239–252, 1999.

Reisinger, Joseph and Mooney, Raymond J. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California, USA, June 2010.

Roque, Francisco S; Slaughter, Laura, and Tkatšenko, Alexandr. A comparison of several key information visualization systems for secondary use of electronic health record content. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 76–83, Los Angeles, California, USA, 2010. Association for Computational Linguistics.

Russell, Stuart J. and Norvig, Peter. *Artificial Intelligence - A Modern Approach (Second Edition)*. Pearson Education, 2005.

Sahlgren, Magnus. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 2006.

Sahlgren, Magnus and Karlgren, Jussi. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11 (03):327–341, 2005.

Sahlgren, Magnus and Karlgren, Jussi. Terminology mining in social media. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 405–414, Hong Kong, China, 2009. Association for Computing Machinery.

Sahlgren, Magnus and Swanberg, David. *Vector based semantic analysis: Modeling linguistic knowledge in computer systems*. PhD thesis, Stockholm University, 2000.

Sahlgren, Magnus; Holst, Anders, and Kanerva, Pentti. Permutations as a means to encode order in word space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Washington, DC, USA, 2008.

Salton, Gerard; Wong, Anita, and Yang, Chung-Shu. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Sarkar, Kamal; Nasipuri, Mita, and Ghose, Suranjan. Using machine learning for medical document summarization. *International Journal of Database Theory and Application*, 4:31–49, 2011.

Schütze, Hinrich. Automatic word sense discrimination. *Computational Linguistics – Special issue on word sense disambiguation*, 24(1):97–123, 1998.

Smadja, Frank. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177, 1993.

Smith, Edward E and Medin, Douglas L. *Categories and concepts*. Harvard University Press Cambridge, MA, 1981.

Sørby, Inger Dybdahl and Nytrø, Øystein. Does the EPR support the discharge process? A study on physicians' use of clinical information systems during discharge of patients with coronary heart disease. *Journal of Healthcare Information Management*, 34(4):112–119, 2005.

Sparck Jones, Karen. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Steinberger, Josef and Křišťan, M. LSA-based multi-document summarization. In *Proceedings of 8th International PhD Workshop on Systems and Control*, volume 7, Balatonfured, Hungary, 2007. Citeseer.

Suominen, Hanna. *Machine learning and clinical text. supporting health information flow*. PhD thesis, University of Turku, 2009.

Suominen, Hanna; Salanterä, Sanna; Velupillai, Sumithra; Chapman, Wendy W.; Savova, Guergana; Elhadad, Noemie; Pradhan, Sameer; South, Brett R.; Mowery, Danielle L.; Jones, Gareth J. F.; Leveling, Johannes; Kelly, Liadh; Goeuriot, Lorraine; Martinez, David, and Zuccon, Guido. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative (CLEF 2013)*, chapter Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, pages 212–231. Springer Berlin Heidelberg, Germany, Valencia, Spain, 2013.

Suominen, Hanna J. and Salakoski, Tapio I. Supporting communication and decision making in finnish intensive care with language technology. *Journal of Healthcare Engineering*, 1:595–614, 2010.

Turney, Peter D and Pantel, Patrick. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

Tversky, Amos. Features of similarity. *Psychological Review*, 84:327–352, 1977.

Van Vleck, Tielman T; Stein, Daniel M; Stetson, Peter D, and Johnson, Stephen B. Assessing data relevance for automated generation of a clinical summary. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 761–765, Chicago, IL, USA, 2007. American Medical Informatics Association.

VanderStoep, Scott W and Johnson, Deidre D. *Research methods for everyday life: Blending qualitative and quantitative approaches*, volume 32. John Wiley & Sons, 2008.

Vapnik, Vladimir; Golowich, Steven E., and Smola, Alex. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*, volume 9, pages 281–287. MIT Press, Cambridge, Massachusetts, 1997.

Velupillai, Sumithra and Kvist, Maria. Fine-grained certainty level annotations used for coarser-grained e-health scenarios. In *Computational Linguistics and Intelligent Text Processing*, pages 450–461. Springer, 2012.

Weaver, Warren. Translation. *Machine Translation of Languages*, 14:15–23, 1955.

World Health Organization, others. International classification of diseases (ICD). 1983.

Wrenn, Jesse O; Stein, Daniel M; Bakken, Suzanne, and Stetson, Peter D. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*, 17(1):49–53, 2010.

Xu, Hua; Stenner, Shane P; Doan, Son; Johnson, Kevin B; Waitman, Lemuel R, and Denny, Joshua C. MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

Yampolskiy, Roman V. *Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing*, chapter Turing Test as a Defining Feature of AI-Completeness, pages 3–17. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.