ENSEMBLES OF SEMANTIC SPACES

On Combining Models of Distributional Semantics with Applications in Healthcare

Aron Henriksson

ISBN 978-91-7649-302-1
ISSN 1101-8526

Typeset by the author using LaTeX
Printed in Sweden by Publit, Stockholm 2015
Distributor: Department of Computer and Systems Sciences

ABSTRACT

Distributional semantics allows models of linguistic meaning to be derived from observations of language use in large amounts of text. By modeling the meaning of words in semantic (vector) space on the basis of co-occurrence information, distributional semantics permits a quantitative interpretation of (relative) word meaning in an unsupervised setting, i.e., human annotations are not required. The ability to obtain inexpensive word representations in this manner helps to alleviate the bottleneck of fully supervised approaches to natural language processing, especially since models of distributional semantics are data-driven and hence agnostic to both language and domain.

All that is required to obtain distributed word representations is a sizeable corpus; however, the composition of the semantic space is not only affected by the underlying data but also by certain model hyperparameters. While these can be optimized for a specific downstream task, there are currently limitations to the extent the many aspects of semantics can be captured in a single model. This dissertation investigates the possibility of capturing multiple aspects of lexical semantics by adopting the ensemble methodology within a distributional semantic framework to create *ensembles of semantic spaces*. To that end, various strategies for creating the constituent semantic spaces, as well as for combining them, are explored in a number of studies.

The notion of semantic space ensembles is generalizable across languages and domains; however, the use of unsupervised methods is particularly valuable in low-resource settings, in particular when annotated corpora are scarce, as in the domain of Swedish healthcare. The semantic space ensembles are here empirically evaluated for tasks that have promising applications in healthcare. It is shown that semantic space ensembles – created by exploiting various corpora and data types, as well as by adjusting model hyperparameters such as the size of the context window and the strategy for handling word order within the context window – are able to outperform the use of any single constituent model on a range of tasks. The semantic space ensembles are used both directly for k-nearest neighbors retrieval and for semi-supervised machine learning. Applying semantic space ensembles to important medical problems facilitates the secondary use of healthcare data, which, despite its abundance and transformative potential, is grossly underutilized.

SAMMANFATTNING

Distributionell semantik gör det möjligt att utarbeta lingvistiska tolkningsmodeller genom observationer av hur ord används i stora mängder text. Genom att modellera ords betydelser i ett vektorrum på basis av information om deras samförekomst kan man med distributionell semantik få en kvantitativ förståelse av (relativ) ordbetydelse i ett oövervakat sammanhang, dvs. där manuella annoteringar inte behövs. Möjligheten att på ett resurseffektivt sätt erhålla ordrepresentationer bidrar till att undvika den flaskhals som uppstår vid fullständigt övervakade angreppssätt till språkteknologi, inte minst eftersom distributionella semantiska modeller är data-drivna och därför kan tillämpas oavsett språk eller domän.

Allt som krävs för att erhålla distribuerade ordrepresentationer är en stor korpus. Det semantiska rummet påverkas dock inte bara av den underliggande datamängden utan även av modellens hyperparametrar. Dessa kan optimeras för att lösa specifika uppgifter men det finns vissa begränsningar beträffande hur många olika semantiska aspekter som kan fångas i en enda modell. Den här avhandlingen undersöker möjligheten att fånga multipla aspekter av lexikal semantik genom att tillämpa ensemblemetodiken inom ett distributionellt semantiskt ramverk för att skapa ensembler av semantiska rum. För detta ändamål undersöks, i ett antal studier, olika strategier för att skapa de ingående semantiska rummen samt för att kombinera dem.

Idén om ensembler av semantiska rum kan tillämpas på olika språk och domäner; användningen av oövervakade metoder är dock särskilt värdefull i situationer där resurser är begränsade, speciellt när annoterade korpusar är sällsynta, vilket är fallet inom den svenska hälsovårdsdomänen. Ensemblerna av semantiska rum utvärderas i denna avhandling empiriskt i uppgifter som har lovande tillämpningar inom hälsovården. Avhandlingen visar att ensemblerna av semantiska rum – skapade genom att utnyttja olika korpusar och datatyper samt genom att justera hyperparametrar såsom kontextfönstrets storlek och strategin för att hantera ordföljd inom detta kontextfönster – kan leda till bättre resultat på en rad uppgifter jämfört med någon av de ingående modellerna. Ensemblerna av semantiska rum används både direkt för hämtning av de k närmaste grannarna samt för semi-övervakad maskininlärning. Genom att tillämpa ensembler av semantiska rum på viktiga medicinska problem underlättas den sekundära användningen av hälsodata, vilken trots sin stora tillgänglighet och transformativa potential utgör en synnerligen underutnyttjad resurs.

LIST OF PUBLICATIONS

This thesis consists of the following original publications:

I Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, Martin Duneld (2014). *Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces*. **Journal of Biomedical Semantics**, 5:6, pp. 1–25.

II Aron Henriksson, Hercules Dalianis, Stewart Kowalski (2014). *Generating Features for Named Entity Recognition by Learning Prototypes in Semantic Space: The Case of De-Identifying Health Records*. In Proceedings of **International Conference on Bioinformatics and Biomedicine**, pp. 450–457, IEEE.

III Aron Henriksson (2015). *Learning Multiple Distributed Prototypes of Semantic Categories for Named Entity Recognition*. **International Journal of Data Mining and Bioinformatics**, Vol. 13, No. 4, pp. 395–411.

IV Aron Henriksson, Jing Zhao, Henrik Boström, Hercules Dalianis (2015). *Modeling Heterogeneous Clinical Sequence Data in Semantic Space for Adverse Drug Event Detection*. In Proceedings of **International Conference on Data Science and Advanced Analytics**, IEEE.

V Aron Henriksson, Jing Zhao, Henrik Boström, Hercules Dalianis (2015). *Modeling Electronic Health Records in Ensembles of Semantic Spaces for Adverse Drug Event Detection*. In Proceedings of **International Conference on Bioinformatics and Biomedicine**, pp. 343–350, IEEE.

VI Aron Henriksson, Maria Kvist, Hercules Dalianis, Martin Duneld (2015). *Identifying Adverse Drug Event Information in Clinical Notes with Distributional Semantic Representations of Context*. **Journal of Biomedical Informatics**, 57:333-349.

ACKNOWLEDGEMENTS

An objective of doctoral studies is to become an independent researcher. However, one certainly does not start out as one and although one may, over time, become more independent, research is not sustainable without the support, in different ways, of a lot of people. I therefore have many whom I would like to thank and acknowledge for their support during these past few years.

A series of serendipitous events first brought me into contact with Hercules, who has been instrumental in my embarking on this journey and has been supportive in seeing me complete it. Thank you for keeping spirits up during the most stressful periods and for leading a close-knit research group. I am also grateful to my co-supervisor, Martin, who has guided me in the world of research in natural language processing, particularly during my first years as a PhD student. Thank you for being supportive and for understanding the perspective of a PhD student.

I have also had the pleasure of working together with many nice people over the years. Former and current members of our research group have provided opportunities for collaboration, but also a sense of belonging, pleasant lunch company and plenty of fun – thank you all! In particular, I would like to thank Maria for fun collaboration and many fond memories from work trips to, in particular, Zinal and San Diego; Sumithra for encouraging me to pursue research and for fun work trips to Australia and USA; and Mia for being a lively, energetic and encouraging colleague, but also for teaching me about the world of healthcare.

It has been my great fortune to belong to other groups that have contributed to a stimulating work and research environment. Thank you to all the members of the DADEL project and the data science group at DSV. I am grateful for the feedback I have received, both on particular studies and my research at large. Henrik, Jussi and Tony deserve special mentions for generously giving me valuable advice at important milestones during my doctoral studies. Thank you to everyone at the Unit of Information Systems and my fellow PhD students at DSV for contributing to a pleasant work environment. Many thanks also to Karolinska University Hospital for giving us access to data, as well as to SLL IT and DSV IT for assistance with technical issues regarding the data.

The opportunity to travel and attend conferences around the world is a wonderful benefit of being a PhD student, especially for an avid traveler like myself. I

have had the opportunity of doing short research visits: twice at the University of California, San Diego and once at the University of Trento in Italy. Wendy, Brian, Mike and Danielle in San Diego, and Marco, Georgiana, German and Ngiah in Rovereto – thank you for inspiring research visits and for graciously hosting me. All of these trips would not have been possible without generously funded scholarships from Helge Ax:son Johnson Foundation, John Söderberg Scholarship Foundation, Knut and Alice Wallenberg Foundation, National Science Foundation, Swedish National Graduate School of Language Technology and Google. Ultimately, all of this has been made possible thanks to the Swedish Foundation for Strategic Research, who have funded my doctoral studies through the DADEL project.

Last, but in many ways most important, thank you to all my family and friends, both near and far: I am immensely grateful to you. A special thanks to my parents, Barbro and Jan-Erik, for always encouraging me in my endeavors, and to my brothers – Jonatan, Jakob, Andreas – and their families for always being there. Undoubtedly the best imaginable side-effect of my doctoral studies is that it brought my future wife and I together: thank you, Jing, for your love and support, especially for bearing with the added stress caused by my unfortunate combination of last-minute and perfectionist tendencies. You have been instrumental, not only in my work and this thesis but in all aspects of my life!

Aron Henriksson
Washington, D.C., November 2015

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 ANALYZING LANGUAGE WITH COMPUTERS

Language is by far our most effective and natural means of communication. A substantial amount of the data we generate is therefore expressed through human languages and typically stored in a largely unstructured form. With the inexorable digitization of the world, this data has become readily accessible for processing by computers and, with the steady rise in computing power, this has been made possible on an increasingly large scale. However, computers tend to prefer dealing with highly structured data and artificial languages that are unambiguous. The aim of *natural language processing*, then, is to equip computers with the ability to analyze and, ultimately, understand human language. Providing computers with the capacity for human language understanding – or mimicking thereof – is crucial for any claims to *artificial intelligence*.

Early approaches to natural language processing – much like in other subfields of artificial intelligence – relied on the construction of rules by human experts. These have been successful for tasks that can readily be broken down into simple, unambiguous rules. However, for tasks that are less straightforward, we struggle to articulate clear rules. Hand-curated systems that are based on rules are not only cumbersome to create, but also become untenable to maintain as the number of rules grows, as it invariably will for many natural language processing tasks as a

result of the ambiguity and irregularity of language. Such systems moreover tend to exhibit poor generalization properties as they are applied to other datasets and over time.

Most contemporary approaches to natural language processing are instead data-driven and based on statistical methods, wherein the ability to perform a particular task is learned from data. The development of such statistical learning algorithms is the focus of another subfield of artificial intelligence, namely *machine learning*. However, the use of data-driven methods does not necessarily allow one to circumvent the need for human experts: the most common – and the most successful – form of machine learning involves human supervision in one way or another. The need for labeled data, typically in the form of human-annotated corpora, constitutes a considerable bottleneck in the development of natural language processing systems. Creating sufficiently large annotated resources for every language, domain and problem is prohibitively expensive, particularly in highly specialized domains where the annotations need to be provided by domain experts. There are, however, also methods for learning that do not require any supervision; these methods instead attempt to discover some structure of interest in the data. For most tasks, relying solely on fully unsupervised learning is unfortunately not feasible. Semi-supervised learning, by exploiting the advantages of both of these approaches, provides a promising middle ground. In this setting, the learning algorithm has access both to a labeled dataset and an unlabeled dataset, where the unlabeled dataset is often assumed to be orders of magnitude larger than the labaled dataset. Many semi-supervised learning approaches use a combination of supervised and unsupervised learning methods.

1.2 COMBINING MODELS OF MEANING

How to model words is central to most natural language processing tasks. When representing texts for supervised machine learning, it is common to employ shallow representations, such as a bag of words or using binary features that indicate whether a given word is present or absent in a text. Such representations, however, fail to account for the deeper semantics of the words. While there are approaches to modeling semantics that attempt to perform a mapping from text to ontological representations, these presuppose the existence of comprehensive,

updated and domain-specific representations; they also give rise to the non-trivial problem of mapping from potentially noisy and diverse uses of language to predefined concepts.

An alternative, and complementary, approach to dealing with semantics and obtaining word representations is *distributional semantics* (see chapter 2). Models of distributional semantics exploit large corpora to capture the relative meaning of words based on their distribution in different contexts. This approach is attractive in that it, in some sense, renders semantics computable: an estimate of the semantic relatedness between pairs of words can be quantified. While observations of language use are needed, manual annotations are not, in effect making it an unsupervised learning method. Models of distributional semantics have been used with great success in many natural language processing tasks, such as information retrieval, various semantic knowledge tests, text categorization and word sense disambiguation. There is also a growing interest in the application of distributional semantics in the biomedical domain [Cohen and Widdows, 2009]. Partly due to the difficulty of obtaining large amounts of clinical data, however, this particular application (sub-)domain has been less explored [Henriksson, 2013].

To construct a distributional semantic space, certain design decisions need to be made, such as which set of language use observations – that is, which corpus – the semantic space should be constructed from. Another important choice concerns the configuration of hyperparameters that the distributional semantic algorithm should employ. In this regard, one critical choice concerns the size of the context window, which determines the scope of co-occurrence events. Some algorithms moreover allow word order within the context window to be handled in different ways. All of these design decisions are important and have an impact on the resulting semantic space [Sahlgren, 2006; Lapesa *et al.*, 2014; Lapesa and Evert, 2014].

In this dissertation, it is posited that making different design choices when constructing semantic spaces leads to different aspects of semantics being captured, and, moreover, that this can be exploited by combining multiple semantic spaces in an ensemble. Such semantic space ensembles are then assumed to improve the predictive performance on a range of natural language processing tasks. The question which this dissertation sets out to answer is thus:

`How can effective ensembles of semantic spaces be created?`

An effective ensemble is here defined as one that outperforms, in the targeted task and with some predefined performance metric, the use of any constituent model, i.e., in this case, a single semantic space. The answer to this question is of considerable importance since it would allow unsupervised methods for obtaining word representations to be exploited further, thereby facilitating the development of natural language processing systems. Here, various ways of creating semantic space ensembles are put forward and evaluated. When creating ensembles, a number of questions naturally present themselves. In an attempt to answer the overarching question, it is thus broken down into the following related research questions:

RQ 1 How can a set of semantic spaces be created to yield effective semantic space ensembles?

RQ 2 How can a set of semantic spaces be combined to yield effective semantic space ensembles?

Answers to these questions are sought in a number of studies: how the papers contribute to answering the overarching research question is shown in Figure 1.1.

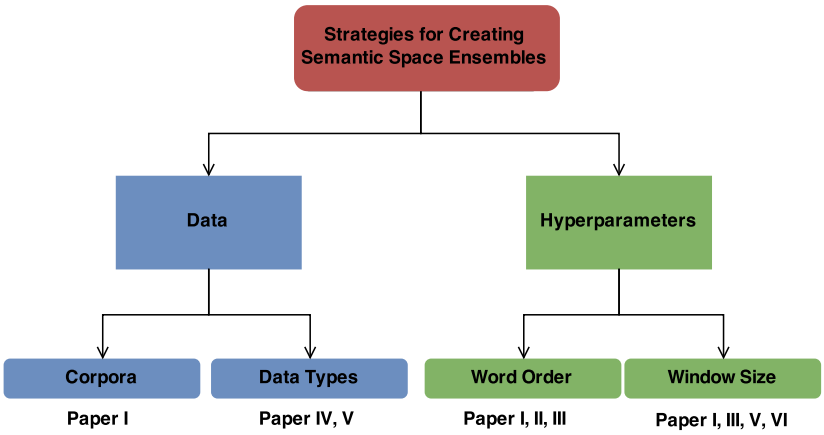

Figure 1.1: OVERVIEW OF PAPERS

1.3 APPLICATIONS IN HEALTHCARE

Although the notion of semantic space ensembles is generalizable across domains and languages, here they are employed in tasks with promising applications in healthcare. In this domain, it is particularly important to be able to capitalize fully on unsupervised methods, as the costs involved in creating annotated resources for supervised machine learning is higher than in many other domains. It is moreover a domain in which the possibility of using machine learning and natural language processing is rather nascent; it has become a promising application area as a result of the the increasing adoption of electronic health records, which, as a positive by-product, gives us access to inexorably growing amounts of healthcare data that can be analyzed by computers. Applying machine learning and natural language processing in this domain is viewed as an important enabler of secondary use of healthcare data, which is currently underutilized, despite its acknowledged transformative potential to improve healthcare, as well as to support medical research and public health activities [Jensen *et al.*, 2012].

There are, however, many challenges involved in learning high-performing predictive models from healthcare, such as the high dimensionality caused by the large number of variables that can be used to describe a given set of observations, as well as the typically ensuing sparsity. There is also an inherent heterogeneity in healthcare data, entailing that the various data types cannot be handled in an identical fashion. The majority of healthcare data is, for instance, expressed in natural language, albeit in a form that is greatly specialized and domain-dependent: clinical text typically does not conform to standard grammar rules and is often littered with shorthand and misspellings [Meystre *et al.*, 2008; Allvin *et al.*, 2011], further exacerbating the aforementioned dimensionality and sparsity issues. There is perhaps, then, a particular need to adapt natural language processing (NLP) techniques to the genre of clinical text, lest the potentially valuable information contained therein should be ignored.

The use of semantic space ensembles is evaluated in the following applications:

SYNONYM EXTRACTION Terminologies that account for variation in language use by linking synonyms and abbreviations to their corresponding concept are important enablers of high-quality information extraction from medical texts. Due to the use of specialized sub-languages in the medical domain, manual construction of semantic resources that accurately reflect language use is both costly and challenging, often resulting in low coverage. Automatic methods that can be used to support the development of such semantic resources are potentially very valuable; distributional semantics can be used to extract candidate synonyms – as well as abbreviated/expanded forms – of terms from large corpora.

NAMED ENTITY RECOGNITION The ability to recognize references to predefined semantic categories of interest – drugs and symptoms, for instance – is an important building-block of natural language processing systems. This task can be automated by learning sequence models from labeled data. Methods that can exploit large amounts of unlabeled data to generate additional features and improve the predictive performance of such systems are, however, of great value. Distributional semantics can be leveraged for creating semi-supervised approaches to named entity recognition.

ADVERSE DRUG EVENT DETECTION Electronic health records are currently being explored as a potential, complementary source for pharmacovigilance, wherein the safety of drugs is monitored to inform decisions on sustained use in the treatment of patients. To that end, it is important to be able to detect and flag for potential adverse drug events on the basis of patient-specific data, not least in order to ensure that the correct diagnosis codes are assigned. One option is to use assigned diagnosis codes as class labels when learning predictive models; another is to learn, from human annotations, to detect and label relations – such as indications and adverse drug events – that hold between mentions of drugs and disorders/symptoms in clinical notes. Distributional semantics can be leveraged to create dense, lower-dimensional representations of the originally high-dimensional and sparse data.

1.4 CONTRIBUTIONS

The contributions in this thesis are all related to the combination of multiple semantic spaces and the use of such ensembles in unsupervised or semi-supervised learning. Specifically, the main contributions are:

1. Multiple corpora from different sub-domains can be exploited to create semantic space ensembles that outperform the use of a single semantic space and a single corpus on a synonym extraction task. This allows out-domain resources to be leveraged in situations where in-domain resources are scarce.
2. Certain hyperparameters of distributional semantic models, such as the size of the context window and the strategy employed for handling word order, can be exploited to create effective semantic space ensembles. This allows different views of the same data to be exploited.
3. Distributional semantics can be extended to non-linguistic sequence data, and semantic space ensembles can be used to exploit heterogeneous data.

1.5 DISPOSITION

The remainder of this thesis is organized into the following parts and chapters:

PART I: PROLOGUE — An extended background to key concepts is provided in this section.

Chapter 2: Distributional Semantics provides a background to distributional semantics, including theoretical foundations, types of models and how different hyperparameters affect the composition of the resulting semantic space.

Chapter 3: Ensemble Models introduces the notion of ensembles and describes why they often outperform single models, as well as strategies for creating effective ensembles.

Chapter 4: Electronic Health Records describes the promise and challenges of applying natural language processing and machine learning in the healthcare domain.

PART II: METHODOLOGY — The research methodology used and related issues are described in this section.

Chapter 5: Methodology describes both the general research strategy, as well as the philosophical assumptions that underpin it. The evaluation framework is laid out in detail, followed by a description of the source of healthcare data that is used in the included studies. Ethical issues are also discussed briefly.

PART III: EMPIRICAL INVESTIGATIONS — The included studies, wherein the use of various semantic space ensembles are evaluated, are in this section summarized application-wise.

Chapter 6: Synonym Extraction investigates the application of semantic space ensembles, created from different corpora and with different word order strategies, to the task of extracting synonyms. A brief description of the main findings is also included.

Chapter 7: Named Entity Recognition investigates the application of semantic space ensembles to named entity recognition. A semi-supervised method that learns prototypical representations in semantic space is proposed and evaluated. A brief description of the main findings is also included.

Chapter 8: Adverse Drug Event Detection investigates the application of semantic space ensembles to adverse drug event detection. In one approach, heterogeneous clinical data is modeled in semantic space ensembles to create representations of healthcare episodes for supervised learning. In another approach, a corpus of clinical notes is annotated, from which relations between named entities are detected and labeled; to that end, the use of distributional semantic representations of context is studied. A brief description of the main findings is also included.

PART IV: EPILOGUE — A broader discussion of the work presented in the studies is embarked upon in this section, including implications and future directions.

Chapter 9: Discussion summarizes and discusses the proposed notion of semantic space ensembles, strategies for creating and combining semantic spaces, as well as their use in important healthcare applications.

Chapter 10: Conclusion provides a recapitulation of the work, highlights main contributions and suggests future directions.

PART I

PROLOGUE

CHAPTER 2

DISTRIBUTIONAL SEMANTICS

For computers to acquire any capacity for human language understanding – remember, the ultimate goal of natural language processing – they first need to be able to account for semantics, i.e., aspects that concern the meaning of linguistic units of various granularity. Endowing computers with semantic knowledge is the focus of computational semantics, a discipline that is concerned with meaning representations that enable subsequent reasoning and computation. An approach to the computational handling of semantics that has long been attractive is based on formal semantics, wherein the meanings of linguistic units are represented in terms of set-theoretic models. These are ideal for capturing the intuition underlying formal semantics, namely that the world is full of objects that have properties, and that relations hold between objects [Clark, 2012]. Formal semantics, however, has very little to say about the meaning of words, which are treated as atomic units.

There is another, complementary, approach that instead exploits the inherent distributional structure of language to acquire the meaning of words. This family of models falls under the umbrella of distributional semantics. Distributional semantic models were initially introduced to overcome the inability to account for the variability of language use in information retrieval [Salton *et al.*, 1975]. In this chapter, the theoretical foundation underlying distributional semantics is described, followed by an introduction to some models of distributional semantics. Finally, model hyperparameters and their impact on the resulting semantic space are described in depth. Much of this is based on [Henriksson, 2013].

2.1 THEORETICAL FOUNDATION

Distributional approaches to meaning make certain assumptions about the nature of language in general and semantics in particular [Sahlgren, 2008]. These assumptions are embedded – explicitly and implicitly – in the distributional hypothesis, which has its foundation in the work of Zellig Harris [1954] and states that words with similar meanings tend to have similar distributions in language[1]. There is thus an assumption that there is a correlation between distributional similarity and semantic similarity: by exploiting the observables to which we have direct access in a corpus – the words and their relative frequency distributions over contexts (distributional similarity) – we can infer that which is hidden, or latent, namely the relative meaning of those words (semantic similarity). For instance, if two words, w_1 and w_2, share distributional properties – in the sense that they appear in similar contexts – we can infer that they are semantically similar along one or more dimensions of meaning.

This rather sweeping approach to meaning presents a point of criticism that is sometimes raised against distributional approaches: the distributional hypothesis does not make any claims about the nature of the semantic (similarity) relation, only to what extent two words are semantically similar. Since a large number of distributionally similar words have potentially very different semantic relations, distributional methods cannot readily distinguish between, for instance, synonymy, hypernymy, hyponymy, co-hyponymy and, in fact, even antonymy[2]. As we will discover, the types of semantic relations that are modeled depend on the employed context definition and thereby also on what type of distributional information is collected [Sahlgren, 2006]. The broad notion of semantic similarity is, however, still meaningful and represents a psychologically intuitive concept [Sahlgren, 2008], even if the ability to distinguish between different types of semantic relations is often desirable.

If we, for a moment, consider the underlying philosophy of language and linguistic theory, the distributional hypothesis – and the models that build upon it – can be

[1] Harris, however, speaks of differences in distributions and meaning rather than similarities: two terms that have different distributions to a larger degree than another pair of terms are also more likely to differ more in meaning. In the distributional hypothesis this is flipped around by emphasizing similarities rather than differences [Sahlgren, 2008].

[2] Since antonyms are perhaps not intuitively seen as semantically similar, some prefer to speak of semantic relatedness rather than semantic similarity [Turney and Pantel, 2010].

characterized as subscribing to a structuralist, functionalist and descriptive view of language. As Sahlgren [2006, 2008] points out, Harris' differential view of meaning can be traced back to the structuralism of Saussure, in which the primary interest lies in the structure of language – seen as a system (*la langue*) – rather than its possible uses (*la parole*). In this language system, signs[3] are defined by their functional differences, which resonates with Harris' notion of meaning differences: the meaning of a given linguistic entity can only be determined in relation to other linguistic entities. Along somewhat similar lines, Wittgenstein [1958] propounded the notion that has been aphoristically captured in the famous dictum: *meaning is use*. The key to uncovering the meaning of a word is to look at its communicative function in the language. The Firthian [Firth, 1957] emphasis on the context-dependent nature of meaning is also perfectly in line with the distributional view of language: *you shall know a word by the company it keeps*. This view of semantics is wholly descriptive – as opposed to prescriptive – in that very few assumptions are made a priori: meaning is simply determined through observations of actual language use. Approaching semantics in this way comes with many other advantages for natural language processing, as it has spawned a family of data-driven, unsupervised methods that do not require hand-annotated data in the learning process.

There are numerous models of distributional semantics that exploit the distributional hypothesis to derive lexical semantic representations from observations of language use in corpora. These can be categorized in various ways. For instance, a distinction is sometimes made between probabilistic models and spatial models [Cohen and Widdows, 2009]. Probabilistic models are today often referred to as topic models [Blei, 2012]; these models tend to use documents as contexts, which are treated as a mixture of topics. Words are represented according to the probability of their occurrence during the discussion of each topic: two words that share similar topic distributions are assumed to be semantically related. Probabilistic models are generative, which means that they can be used to estimate the likelihood of unseen documents. Spatial models of distributional semantics – which are the only ones considered in this thesis – represent words in vector space, where proximity indicates semantic similarity: words that are close to each other in vector space are assumed to be close in meaning; words that are distant are semantically unrelated. Running an algorithm that implements a spatial model

[3] In structuralism, a linguistic sign is composed of two parts: a signifier (a word) and a signified (the meaning of that word).

over a corpus results in a semantic space[4]. In spatial models, proximity can be measured in several ways; however, the most common way is to calculate the cosine of the angle between two vectors, **u** and **v**, with length n as follows:

$$\cos(\Theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (v_i)^2}} \tag{2.1}$$

A higher cosine similarity score indicates a higher degree of semantic similarity. The notion of modeling linguistic meaning in vector space is sometimes referred to as the geometric metaphor of meaning and intuitively captures the idea that terms can, metaphorically speaking, be near in meaning, as well as the fact that the meaning of words can shift over time and across domains [Widdows, 2004]. For these reasons, spatial models are by many considered realistic and intuitive from a psychological perspective [Landauer and Dumais, 1997].

Another distinction is made between context-counting models and context-predicting models [Baroni *et al.*, 2014]. Early models of distributional semantics were essentially context-counting in the sense that they relied fundamentally on counting the contexts – typically co-occurrences with words in a window around the target word – in which a word appears. Since then, alternative models have been proposed that rely on learning to predict the contexts in which a word appears. Most of these models have been inspired by the success of deep learning methods. In several systematic evaluations, context-predicting models have been shown to outperform context-counting models [Baroni *et al.*, 2014] and to produce robust results across many semantic tasks [Schnabel *et al.*, 2015]. Although this distinction is pertinent, it has recently been shown, both theoretically and empirically, that there is no qualitative difference between these two types of models; they simply offer different computational means of arriving at the same type of model [Levy and Goldberg, 2014; Pennington *et al.*, 2014]. For instance, the context-predicting skip-gram model (see section 2.3) is implicitly factorizing a PMI[5]-weighted word-context matrix shifted by a global

[4] Several alternative terms exist in the literature, including word space and distributional semantic space.

[5] Pointwise Mutual Information (see section 2.2)

constant [Levy and Goldberg, 2014], while the performance gains are essentially a result of hyperparameter optimizations rather than any substantial differences in the model [Levy *et al.*, 2015].

2.2 CONTEXT-COUNTING MODELS

Context-counting models of distributional semantics (see [Turney and Pantel, 2010] or [Clark, 2012] for a more elaborate introduction) – as the name suggests – essentially rely on counting the contexts in which words appear. Differences between various context-counting models concern the manner in which the word representations are obtained. These are generally derived from a word-context matrix[6] M where each row i corresponds to a word and each column j refers to a context in which the word appears; each matrix entry M_{ij} corresponds to the frequency with which a word occurs in the given context. The meaning of a word is hence represented as a frequency distribution over contexts. The context definition may vary both across and within models. Documents typically define the context in probabilistic models. In spatial models, however, using words as context elements is the most prevalent approach. In context-counting models, word co-occurrences are counted within a window surrounding each instance of a word[7]. There are also linguistically motivated context definitions, achieved with the aid of, for instance, a syntactic parser, where the context is defined in terms of syntactic dependency relations [Padó and Lapata, 2007]. An advantage of employing a more naïve context definition is, however, that the language-agnostic property of distributional semantics is thereby retained: all that is needed is a tokenized corpus.

Using raw counts of co-occurrences with all other words in the vocabulary is problematic for several reasons. On the one hand, it fails to account for the fact that not all co-occurrence events are equally significant from a semantic point of view. For instance, the fact that a word has co-occurred with a function word like

[6]There are also pair-pattern matrices, where rows correspond to pairs of words and columns correspond to the patterns in which the pairs co-occur. These have been used for identifying semantically similar patterns [Lin and Pantel, 2001] – by comparing column vectors – and analogy relations [Turney and Littman, 2005] – by comparing row vectors. Pattern similarity is a useful notion in paraphrase identification [Lin and Pantel, 2001]. It is also possible to employ higher-order tensors (vectors are first-order tensors and matrices are second-order tensors), see for instance [Turney, 2007].

[7]The window is typically symmetric around the target word and the size of the window is typically static and predefined.

the, which, depending on the precise context definition, co-occurs with most other words, contributes very little to determining the meaning of that word. As a result, it is common to perform some form of weighting to account for this observation. Several different weighting strategies exist; however a common strategy is to use (positive[8]) pointwise mutual information (PMI) weighting [Church and Hanks, 1990]:

$$PMI(x,y) = log\frac{P(x,y)}{P(x)P(y)}, \tag{2.2}$$

where x and y are words. PMI compares the joint probability of x and y, i.e., the probability of observing the words together, with the probabilities of observing x and y independently. Another problem has to do with high dimensionality and the fact that word co-occurrences are rare events: most words – with the exception of, for instance, function words – only co-occur with a small subset of other words. Regardless of which context definition is employed, but particularly so when words define the context, the number of contexts is exceptionally large – as large as the size of the vocabulary. Working directly with such high-dimensional – where the dimensionality is equal to the number of contexts – and inherently sparse data involves unnecessary computational complexity and comes with significant memory overhead, in particular since most cells in the matrix would be zero as a result of the rarity of word co-occurrences. Data sparsity as a result of high dimensionality is commonly referred to as the curse of dimensionality. One remedy is to perform some form of feature selection. A naïve feature selection strategy is simply to select the K most frequent words, which could be proportionate to the size of vocabulary. Another solution is to project the high-dimensional data into a lower-dimensional space, while approximately preserving the relative distances between data points. The benefit of dimensionality reduction is two-fold: on the one hand, it reduces complexity and data sparsity; on the other hand, it has also been shown to improve the coverage and accuracy of term-term associations when employing a document-level context [Landauer and Dumais, 1997], as, in this reduced semantic space, terms that do not necessarily co-occur directly in the same contexts will nevertheless be clustered about the same subspace, as long as they appear in similar contexts, i.e. have neighbors in common (co-occur with the same terms). In this way,

[8]Positive pointwise mutual information is a variant that simply discards non-positive values [Niwa and Nitta, 1994].

the reduced space can be said to capture higher order co-occurrence relations and latent semantic features.

There are many different dimensionality reduction techniques, e.g., singular value decomposition (SVD)[9]. Singular value decomposition decomposes the term-context matrix into a reduced-dimensional matrix of a predefined dimensionality d that best captures the variance between data points in the original matrix, given a particular d. Dimensionality techniques like SVD can, however, be computationally expensive. As a result, more efficient models of distributional semantics were proposed that did not need to rely on such techniques.

Random indexing [Kanerva *et al.*, 2000] is an incremental, scalable and computationally efficient context-counting model of distributional semantics, but it is also a dimensionality reduction technique. However, in one sense, explicit dimensionality reduction can be said to be circumvented while achieving essentially the same effect: instead of constructing an initial term-context matrix, which is subsequently reduced, a lower-dimensional semantic space is constructed in an incremental fashion. The dimensionality of the semantic space is chosen a priori as a model hyperparameter. The trick is to represent contexts, not with orthogonal one-of-K representations[10] – where K is the number of contexts, potentially the size of the vocabulary – but with nearly orthogonal d-dimensional sparse vectors in which a small set of 1s and -1s[11] have been randomly distributed, with the rest of the elements set to zero. By generating sparse vectors of a sufficiently large dimensionality in this way, they will, with a high probability, be nearly orthogonal. Random indexing – like random projection [Papadimitriou *et al.*, 1998] and random mapping [Kaski, 1998] – is motivated by the Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1984]. This asserts that the relative distances between points in a high-dimensional Euclidean (vector) space will be approximately preserved if they are randomly projected into a lower-dimensional subspace of sufficiently high dimensionality. This is what random indexing exploits by representing context elements as nearly orthogonal vectors that are randomly projected into a reduced-dimensional space: since there

[9] In one of the early context-counting models, latent semantic indexing/analysis (LSI/LSA) [Deerwester *et al.*, 1990; Landauer and Dumais, 1997], SVD was used for dimensionality reduction.

[10] The one-of-K representation, or one-hot encoding, means that only one vector component is 1 and all other vector components are 0.

[11] These are thus ternary vectors that allow three possible values: 1s, 0s and -1s. Allowing negative vector elements ensures that the entire vector space is utilized [Karlgren *et al.*, 2008]. The proportion of non-zero elements is typically 1-2%.

are more nearly orthogonal than truly orthogonal vectors in a high-dimensional space [Kaski, 1998], randomly generating sparse vectors is a good approximation of orthogonality. In the construction of a semantic space with random indexing and using words to represent context, each unique word $w \in W$ is assigned an index vector $\mathbf{i}_w$ and a semantic vector $\mathbf{s}_w$ of the same dimensionality. The index vectors represent context elements and are only used for semantic space construction. The meaning of a word w, represented by $\mathbf{s}_w$, is then the sum[12] of all the index vectors of the words with which w co-occurs within a window of a certain pre-defined size.

Making the dimensionality a model hyperparameter gives the method attractive scalability properties since the dimensionality will not grow with the size of the data, or, more specifically, the number of contexts[13]. Moreover, the incremental approach of random indexing allows new data to be added at any given time without having to rebuild the semantic space: if a hitherto unseen word is encountered, it is simply assigned a random index vector and a zero semantic vector. Its ability to handle streaming data makes it possible to study real-time acquisition of semantic knowledge [Cohen and Widdows, 2009].

2.3 CONTEXT-PREDICTING MODELS

Context-predicting models of distributional semantics rely not on explicitly counting the contexts in which words appear, but on training predictive models in such a manner that meaningful word representations – or word embeddings[14] as they are often called in this setting – can be learned. A context-predicting model $W : \mathit{words} \rightarrow \mathbb{R}_d$ is a parameterized function mapping words to dense[15] d-dimensional vectors. The function is typically a lookup table, parameterized by a matrix θ, with a row for each word $W_\theta(w_n) = \theta_d$. W is initialized to have random vectors for each word and then learns to have meaningful vectors in order to perform some task. The precise prediction task varies somewhat

[12] Sometimes weighting is applied to the index vector before it is added to the semantic vector, for instance according to the distance and/or direction of the co-occurring word to the target word.

[13] The dimensionality may, however, not remain appropriate irrespective of data size [Henriksson and Hassel, 2013]

[14] In contrast, they are often called distributional vectors or distributed vectors when using context-counting models.

[15] Context-counting models, on the other hand, typically produce sparse vectors, i.e., when dimensionality reduction is not performed.

across models; the learning task itself is, however, not the end goal but merely constructed such that meaningful word representations can be obtained. From the point of view of distributional semantics – following the distributional hypothesis – the learning task should be constructed such that we obtain similar word representations for words with similar meanings. While the word representations that are learned will be useful for performing the constructed task, they are typically intended to be employed for other tasks. It should also be pointed out that context-predicting models are typically spatial models, i.e., words are represented in vector space and the semantic similarity between two words can be quantified using a distance/similarity measure like cosine similarity.

Context-predicting models have come out of research in deep learning[16] and neutral network-based language models [Bengio *et al.*, 2003]. In order to obtain meaningful word representations, deep learning is however not necessary; in fact, most successful models are based on training shallow neural networks with a single hidden layer. The weights learned in the hidden layer give us W. It is also worth noting that context-predicting models employ supervised learning algorithms but do so in a fully unsupervised setting: the examples are obtained from an unlabeled corpus.

There are several context-counting models of distributional semantics. One of the most successful ones is implemented in word2vec [Mikolov *et al.*, 2013a,b]. Two different architectures have been proposed that differ in the construction of the learning task: continuous bag of words (CBOW) and skip-gram (Figure 2.1).

The learning task in CBOW is to predict the word given the context. In the skip-gram model, the task is instead to predict the context given the word. The training objective of the skip-gram model is to maximize the dot product between the vectors of frequent word-context pairs and to minimize it for random word-context pairs [Levy and Goldberg, 2014]. Given a set D of words i and their contexts c, the objective function of the model is to set the parameters Θ that maximize $p(c|i;\Theta)$ [Goldberg and Levy, 2014]:

$$\arg\max_{\Theta} \prod_{(i,c)\in D} p(c|i;\Theta)$$

[16]The idea of distributed representations for symbols can be traced back to the 1980s [Hinton, 1986], while the idea of word embeddings was developed before the revival of deep learning in 2006.

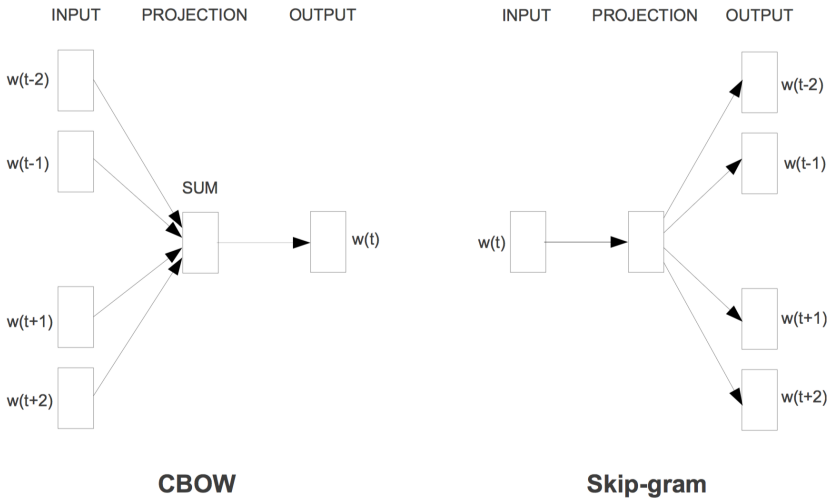

Figure 2.1: TWO ARCHITECTURES IMPLEMENTED IN WORD2VEC; ILLUSTRATION TAKEN FROM [MIKOLOV *et al.*, 2013]

The context, as in context-counting models, is typically defined as an adjacent item within a symmetric window of a pre-specified size around the input item. Both CBOW and skip-gram are efficient to train; however, skip-gram tends to result in better word representations for infrequent words, while CBOW is somewhat faster and more suitable for very large corpora [Mikolov *et al.*, 2013a].

2.4 MODEL HYPERPARAMETERS

There are two main factors that influence the structure of the semantic space: (1) the underlying data and (2) various hyperparameters of the employed model. Different models of distributional semantics have different hyperparamters; however, one that is common to all is the definition of context. How context is defined has an instrumental role in determining which aspects of semantics, in particular which semantic relations, are captured in the semantic space. One important distinction exists that has been highlighted is that between syntagmatic and paradigmatic relations; which one is modeled depends on the context

definition that is employed [Sahlgren, 2006]. It has been shown that a context definition based on documents captures syntagmatic relations between words; these are words that tend to co-occur (e.g., {car, engine, road}). When employing a context definition based on words, syntagmatic relations between words are typically captured in the semantic space; these are words that do not themselves co-occur but share neighbors (e.g., {car, automobile}). Synonyms are hence prime examples of syntagmatic relations.

More recently it has been shown that both paradigmatic and syntagmatic relations can, in fact, be captured by models that define context based on co-occurrences with words within a (sliding) window of a certain size [Lapesa *et al.*, 2014]. Here, the size of the context window plays an important in contrasting these broad categories of relations: using a narrow window size tends to better capture paradigmatic relations, while using a somewhat larger window size tends to better capture syntagmatic relations. That the size of the context window does indeed matter has been confirmed in several studies [Peirsman *et al.*, 2008; Kiela and Clark, 2014]; moreover, the optimal window-size tends to be dependent on the task [Lapesa and Evert, 2014] as well as the size (and noise) of the underlying corpus [Kiela and Clark, 2014].

Models of distributional semantics typically treat the context window as a bag of words. As a result, these models have received some criticism for failing to account for word order and their role in determining word meaning. There are, however, models that have attempt to take into account word order information. For instance, variants[17] of random indexing take into account such information in an efficient manner [Sahlgren *et al.*, 2008]. This is achieved by simply shifting the vector components of the random index vector according to the corresponding word's distance and/or direction in relation to the target word. When shifting the vector components according to direction alone, the resulting semantic vectors are sometimes referred to as direction vectors, while, when shifting the vector components according to direction and distance, the resulting semantic vectors are sometimes referred to as order vectors. For instance, before adding the index vector of a word that occurs two positions to the left of the target word, the vector components of the index vector are shifted two positions to the left in the case of order vectors and one position to the left in the case of direction vectors. Each word thus defines multiple contexts, resulting – in the case of order vectors –

[17] These are inspired by the work of [Jones and Mewhort, 2007] and their BEAGLE model that uses vector convolution to incorporate word order information in semantic space.

in one index vector for each possible position relative to the target word in the context window[18]. However, performing this procedure on the fly before adding an index vector to a semantic vector means that only one index vector per context element needs to be stored in memory. Incorporating word order information in models of distributional semantics constrains the types of semantic relations that are captured. It has, for instance, been shown to improve results on several tasks [Sahlgren *et al.*, 2008; Henriksson *et al.*, 2013a].

There of course many other model hyperparameters, including the dimensionality of the semantic space, which context elements to consider[19] and various strategies for weighting the co-occurrence events. These are, however, not considered in this thesis and therefore not elaborated on. The purpose of this section is rather to describe the importance of certain model hyperparameters on the structure of the resulting semantic space. While these hyperparameters can be optimized for a particular task – i.e., a certain dataset – it is generally not possible to select a set of hyperparameters that is optimal across tasks: the order of models depends on the evaluation criteria that are used [Schnabel *et al.*, 2015].

[18]This, of course, affects the near-orthogonality property of random indexing; theoretically, the dimensionality needs to be higher when employing a variant of random indexing that takes into account word order information.

[19]This includes corpus preprocessing, e.g., stemming and lemmatization. This is sometimes referred to as feature granularity and has been shown to lead to improvements [Kiela and Clark, 2014].

CHAPTER 3

ENSEMBLE MODELS

In the context of predictive modeling, ensembles are models that are composed of a set of individual models whose predictions are combined. Research has shown that ensembles often perform better than any of the individual models, or base models, of which they are composed [Opitz and Maclin, 1999; Dietterich, 2000a]. There is a wide range of ensemble models that employ different strategies for creating the base models and for combining their predictions. These two fundamental components of ensembles – (1) creation strategies and (2) combination strategies – are also investigated in this thesis, where, in contrast to previous work, the ensembles are composed of distributional semantic spaces. In fact, ensembles have primarily been investigated in a fully supervised setting; however, the notion of ensembles has also been adopted in unsupervised and semi-supervised settings [Okun and Valentini, 2008], which are the focus of this thesis.

In this chapter, key concepts are introduced in a brief summary of the theoretical foundation underlying ensemble models. This is followed by an overview of commonly employed creation and combination strategies, respectively. Finally, the related area of information fusion is introduced.

3.1 THEORETICAL FOUNDATION

In supervised learning, the learning algorithm is given a training sample S with examples of the form $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ for some unknown target function $f : \mathcal{X} \rightarrow \mathcal{Y}$. The learning algorithm uses the training examples to choose a final hypothesis g from a set of hypotheses $\mathcal{H} = \{h_1, ..., h_j\}$ such that $g \approx f$. The $\mathbf{x}_i$ values are typically vectors of the form $\langle x_{i,1}, x_{i,2}, ..., x_{i,m} \rangle$ whose components are discrete or real-valued and are referred to as features of $\mathbf{x}_i$. The y values are, in the case of classification, a discrete set of classes $c \in dom(y)$, or, in the case of regression, drawn from the real line.

An ensemble of classifiers is, then, a set of hypotheses (or base models) whose individual predictions are combined in some way to classify new examples. The reason for constructing ensembles is that they have been shown often to result in higher predictive performance than the base models of which they are composed. Much research has gone into devising strategies for constructing effective ensembles [Löfström, 2015] and it has been shown that this essentially requires the base models to be both accurate and diverse [Hansen and Salamon, 1990]. An accurate base model is one that has an error rate that is lower than random guessing, while two base models are diverse if they make different errors of new data points.

These conditions for creating effective ensembles can be traced as far back as the 18th century and the works of Marquis de Condorcet [1785]. Condorcet formulated a theorem, known as Condorcet's jury theorem [Austen-Smith and Banks, 1996], which states that the error of the majority of a jury decreases with the number of jury members. This theorem holds under the assumption that each member is more likely to be correct than wrong (i.e., are accurate), but also requires that the members make the errors independently (i.e., are diverse). The latter means, for example, that nothing is gained from forming a jury whose members always agree; the overall error will be no lower than the error of each single member. In the framework of ensemble learning, each base model in the ensemble thus corresponds to a jury member. Besides the number of base models in the ensemble, there are hence two components that affect the predictive performance: the performance of each base model and to what extent the base models vary in their predictions. The latter is often referred to as the diversity of the ensemble [Kuncheva and Whitaker, 2003]. In a regression framework, the (squared) error E of the ensemble is directly related to the average (squared)

error A of the ensemble members, and their diversity D, i.e., the average (squared) deviation of each single prediction from the ensemble prediction, as shown by the following equation [Krogh and Vedelsby, 1995]:

$$E = \bar{A} - \bar{D} \tag{3.1}$$

The above states that the ensemble error can be no higher than the average model error, and that the more diversity there is, the lower the ensemble error will be. It should, however, be noted that using the above directly in the search for an optimal ensemble is not straightforward, as there is normally a strong interplay between diversity and average base model performance: perfect models will, for instance, agree on all predictions. When it comes to classification, there is unfortunately no straightforward decomposition of ensemble performance into average base model performance and diversity. However, a large number of alternative diversity measures have been proposed in the literature [Kuncheva and Whitaker, 2003], although their connection to ensemble performance has been shown to be questionable. Many of these are pairwise measures that compare the predictions of a pair of base models, allowing one to obtain an estimate of the ensemble's overall diversity. One such measure is Yule's Q-statistic [Yule, 1900]:

$$Q = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \tag{3.2}$$

where, for a pair of classifiers, N_{00} denotes the number of cases where both are incorrect, N_{11} the number of cases where both are correct, N_{01} the number of cases where the first is incorrect but not the other, and finally N_{10} where the first is correct and the second incorrect. Q is between -1 and 1: classifiers that tend to recognize the same objects correctly will have positive values of Q, and those that commit errors on different objects – indicating diversity – will render Q negative.

To estimate the diversity in ensembles of semantic spaces, the use of the Q-statistic is only applicable when the ensembles are used for providing features to a classifier. A common way to use and gain insight into semantic spaces is to explore local neighborhoods by inspecting the k nearest neighbors of given items. By pairwise comparing the nearest neighbors of the same items across semantic spaces, one can obtain some estimate of their diversity. This essentially entails comparing ranked lists, for which many methods exist. These can be categorized

into rank correlation methods and set-based methods [Webber *et al.*, 2010]. Kendall's Tau [Kendall, 1938] belongs to the former and essentially measures the probability of two items being in the same order in the two ranked lists:

$$\tau = \frac{C - D}{N}, \tag{3.3}$$

where C is the number of concordant pairs, i.e., the number of pairs for which the relative ordering is preserved in the two lists; D is the number of discordant pairs, i.e., the number of pairs for which the relative ordering is reversed; and N is the total number of pairs, $\frac{n(n-1)}{2}$, from a list with n items. The coefficient must be in the range $[-1, 1]$, where a value of 1 indicates perfect agreement between the two rankings, a value of -1 indicates perfect disagreement between the two rankings, and a value of 0 indicates that the two rankings are independent. A problem with Kendall's Tau is, however, that it is unweighted, which means that the rank position of an item has no effect on the final similarity score. The property that is often desired is known as top-weightedness. Set-based metrics exist that satisfy the top-weightedness criterion. The basic idea is to calculate the fraction of content overlapping at different depths and then to return the average overlap. For two ranked sets, A and B, the average overlap score o between them can be defined as follows:

$$o = \frac{\sum_{i=1}^{N} (|\{A_1, \ldots, A_i\} \cap \{B_1, \ldots, B_i\}|/i)}{N}, \tag{3.4}$$

where N is the length of both sets and the o coefficient must be in the range $[0, 1]$. This approach is naturally top-weighted, i.e., it gives more importance to items that are ranked highly, since observing a common item at a higher rank position contributes to all the lower-ranked intersections.

Additional explanations for the power of ensembles are provided by Dietterich [2000a], from three points of view: statistical, computational and representational. The statistical reason arises when the size of the training data is insufficient in relation to the size of the hypothesis space $\mathcal{H}$. Without sufficient data, the learning algorithm may find many different hypotheses in $\mathcal{H}$ that obtain the same performance on the training data. By creating an ensemble made up of these accurate classifiers, the risk of selecting the wrong base model is reduced. That

is, by averaging the set of hypotheses $h_1, ..., h_j$ that collectively make up our ensemble, we are more likely to end up with a final hypothesis g that better approximates our unknown target function f. The second reason is computational, which explains the power of ensembles even in situations where there is a sufficient amount of training data. The computational explanation is that many learning algorithms may get stuck in local optima when searching for g in $\mathcal{H}$ using, e.g., gradient descent in the case of neural networks or greedy splitting rules in the case of decision trees. An ensemble can instead be created that effectively runs the local search from multiple different starting points, thereby reducing the risk of choosing one that gets stuck in a local optimum. The third reason is representational, which concerns the hypothesis space $\mathcal{H}$ that a learning algorithm is restricted by: many times the true function f cannot be represented by the set of hypotheses in $\mathcal{H}$. An ensemble may be better able to approximate f by forming weighted sums of the available hypotheses and, in effect, expanding $\mathcal{H}$[1].

3.2 CREATION STRATEGIES

There are many strategies for creating ensembles. The base models can be trained from a single data set, different versions of the same data set, or even different data sets that may have been collected from different sources. One of the most common strategies for generating multiple hypotheses, i.e., multiple base models, is based on manipulation of the training sample S as a means of achieving diversity. The learning algorithm is hence run several times on different versions of S. This strategy has been shown to work particularly well in conjunction with unstable learning algorithms, i.e., algorithms that lead to very different models as a result of small changes in S [Dietterich, 2000a]. One popular method for manipulating the training examples is bagging [Breiman, 1996], which entails creating different versions of S by means of resampling. Each base model is trained on a bootstrap replicate of S, i.e., m training examples are drawn randomly with replacement from the original training set of m items. Each bootstrap replicate contains, on average, 63.2% of the original training set, with several training examples appearing multiple times. There are also other ways of manipulating the training

[1] It should be noted that many learning algorithms, e.g., neural networks and decision trees, are highly flexible and, in principle, may explore all possible hypotheses. However, given a finite amount of training data, only a finite number of hypotheses will be explored and the search procedure will be discontinued once an hypothesis that fits the training data has been found.

data to obtain diverse base models, e.g., cross-validated committees [Parmanto *et al.*, 1996]. Other means of obtaining diverse base models is to manipulate, for instance, the input features or the output targets. By only providing a subset, possibly randomly selected, to the learning algorithm at each run, multiple and potentially diverse base models can be obtained. The random forest learning algorithm [Breiman, 2001], which is used for generating predictive models in several of papers included in this thesis, makes use of both bagging and random feature subset selection to generate diverse base models. Random forest constructs an ensemble of decision trees, which together – through majority voting – decide which class label to assign to a new example. Each tree in the forest is built by creating a bootstrap replicate of the training examples, while a subset of the features are randomly sampled when splitting at each node in the tree-building procedure.

Yet another popular ensemble creation strategy is known as boosting [Schapire, 1990]. In contrast to bagging, which allows for training of base models independently of each and in parallel, boosting is an inherently sequential procedure that results in a series of base models, where the training set used for training each model is based on the performance of the previous classifier(s): training examples that are predicted incorrectly by the previous models are selected more often, or given more weight, than training examples that were predicted correctly. Boosting hence attempts to generate models that perform better on examples for which the current ensemble is predicting incorrectly.

Another general strategy for creating base models for an ensemble is to manipulate the model hyperparameters, i.e., using a different set of hyperparemeters when training each base model. Many times this can be viewed as essentially injecting randomness into the learning algorithm. When training neural networks using backpropagation, for instance, the initial weights are set randomly; using different initial weights may result in very different models [Pollack, 1990]. Randomly setting hyperparemeters has also been successfully applied to create ensembles of decision trees [Dietterich, 2000b]. Ensembles can also be created by applying different learning algorithms to the same data.

3.3 COMBINATION STRATEGIES

Once a set of base models have been created, one must decide on a strategy for combining their output. This of course depends on whether the combination is performed in a classification or regression setting. Combination strategies can generally be divided into whether they are supervised and unsupervised, i.e. where the combination strategy is learned or performed by consensus. There are pros and cons of each. Learning a combiner, for instance, allows useful feedback to be provided by the labeled data and help to improve accuracy; however, labeled data needs to be set aside for this and this approach may moreover lead to overfitting. Combination by consensus, on the other hand, is unsupervised and does not depend on labeled data, which, however, means that assumptions have to be made about how the combination should be performed a priori.

In a classification setting, the following subdivision of combination strategies into three levels has been suggested [Xu *et al.*, 1992]: (1) the abstract level, (2) the rank level and (3) the measurement level. On the abstract level, each base model merely outputs its preferred class label; on the rank level, each base model outputs a ranked list of all class labels according to preference; finally, on the measurement level, each base model outputs a numeric value that quantifies the degree to which the example belongs to each class. The most straightforward combination strategy belongs to the abstract level and is typically referred to as majority voting, which simply assigns the class label that the majority of the base models predicts:

$$class(\mathbf{x}) = \underset{c_i \in dom(y)}{\arg\max} \sum_k f(y_k(\mathbf{x}), c_i) \tag{3.5}$$

where $y_k(\mathbf{x})$ is the prediction of the kth base model and $f(y,c)$ is an indication function, defined as $g(y,c) = 1$ when $y = c$ or 0 when $y \neq c$. In the case of a tie, one could, for instance, randomly select among the tied classes or select the class with the highest prior probability. A variant of this is applicable when using probabilistic classifiers, whereby the class probabilities are taken into account when deciding the ensemble prediction. This strategy hence belongs on the

measurement level. One option is to select the class label with the highest mean probability:

$$class(\mathbf{x}) = \underset{c_i \in dom(y)}{\arg\max} \left(\sum_k \hat{P}_{M_k}(y = c_i|\mathbf{x})/K \right) \tag{3.6}$$

where $\hat{P}_{M_k}(y = c|\mathbf{x})$ denotes the probability of class c given an instance $\mathbf{x}$ and K is the total number of base models. Another option is to select the class label chosen by the base model that is most certain about its prediction, i.e., has the largest difference between the probabilities of two classes:

$$class(\mathbf{x}) = \underset{k \in 1 \ldots K}{\arg\max} \left| \hat{P}_{M_k}(y = c_i|\mathbf{x}) - \hat{P}_{M_k}(y = c_i|\mathbf{x}) \right| \tag{3.7}$$

Another variant is weighted majority voting, where weights are assigned to the base models according to their performance on the training set or, in the case of bagging ensembles, out-of-bag (OOB) examples. One can either weight the probabilities or the predictions based on, for example, the base models' performance on OOB examples. The following strategy assigns a proportional weight to each base model according to its OOB accuracy, and then multiplies the corresponding weight to each model's predicted class probabilities:

$$class(\mathbf{x}) = \underset{c_i \in dom(y)}{\arg\max} \sum_k \left(w_k \times \hat{P}_{M_k}(y = c_i|\mathbf{x}) \right), \tag{3.8}$$

where w_k is the weight assigned to the kth base model, $\hat{P}_{oob_k}$, defined as $w_k = \hat{P}_{oob_k} / \sum_k \hat{P}_{oob_k}$. The following strategy similarly assigns a proportional weights to each base model according to its OOB accuracy, but then weights the votes and classifies an instance according to the class label with the highest weighted votes:

$$class(\mathbf{x}) = \underset{c_i \in dom(y)}{\arg\max} \sum_k \left(w_k \times f(y_k(\mathbf{x}), c_i) \right) \tag{3.9}$$

Many more combination strategies are described in [Rokach, 2010]. Stacking [Wolpert, 1992] is an example of a supervised combination strategy, whereby a

combiner learns which base models are reliable and which are not. Instead of using the original input features, the predictions output by the base models are used as input features.

3.4 INFORMATION FUSION

An area that, in some ways, is related to ensemble learning, and that has benefited from the use of ensemble models, is information fusion, which concerns the modeling of information from different sources and different time points to support decision making [Boström *et al.*, 2007]. There is an abundance of heterogeneous data sources that often tend to complement each other; however, as they typically cannot be handled in a uniform manner, it is a non-trivial task to combine them in an effective way.

A fundamental question in information fusion concerns when to combine the potentially heterogeneous data types. In the context of predictive modeling, a distinction is often made between early and late fusion. Early fusion entails that the combination is done prior to prediction, while late fusion entails that predictions are first made using models trained separately on each data type, independently of each other, and then combined. Early fusion typically corresponds to feature fusion, while late fusion corresponds to classifier fusion. Ensemble models are generally a form of classifier fusion; in the context of information fusion, however, the classifiers are typically trained using data from different sources and that are of potentially different types. Feature fusion generally means that features are generated from each data source and combined in some way. Often this could simply mean that the feature vectors are concatenated prior to learning; however, there are also more sophisticated feature fusion strategies [Yang *et al.*, 2003]. In this thesis, the term feature fusion is limited to feature concatenation. Comparisons of information fusion strategies have been made, see, e.g., [Böstrom, 2007].

CHAPTER 4

ELECTRONIC HEALTH RECORDS

Electronic health records (EHRs) have the primary function of storing data generated in the healthcare process, serving as a crucial archiving and communication tool. A positive by-product of the increasing adoption of EHR systems is that we thereby gain access to inexorably growing amounts of healthcare data that can be processed by computers. This data is potentially very valuable as it contains longitudinal observations of patients and their treatment. The exploitation of EHRs for secondary use is increasingly being acknowledged as having transformative potential to improve healthcare, as well as to support medical research and public health activities [Meystre *et al.*, 2008; Jensen *et al.*, 2012].

There are, however, also many challenges in analyzing EHR data, not least as a result of the issues that arise with high dimensionality and sparsity. The data contained in EHRs is moreover heterogeneous and it is non-trivial how this should be handled, in particular how the various data types can be exploited for predictive modeling. In this chapter, an introduction to the growing area of clinical language processing is first given, followed by backgrounds to the three important applications in which the semantic space ensembles are utilized: synonym extraction, named entity recognition and adverse drug event detection.

4.1 HETEROGENEOUS DATA

Electronic health records contain heterogeneous data. While the types of data are numerous, a crude distinction is sometimes made between structured and unstructured data. Structured EHR data typically includes basic characteristics of the patients and various codes that correspond to, for instance, diagnoses and drugs, as well as various clinical measurements, such as blood tests and physical characteristics. Clinical notes, written or dictated by clinicians in natural language, are, on the other hand, inherently more unstructured.

Clinical notes are written in the clinical setting and denote any narrative that appears in EHRs, ranging from short texts describing a patient's chief complaint to long descriptions of a patient's medical history [Meystre *et al.*, 2008]. The language used in clinical text can also be characterized as a sublanguage with its own particular structure, distinct from both general language and other sublanguages [Friedman *et al.*, 2002]. The noisy character of clinical text makes it much more challenging to analyze computationally than general language text, such as newspaper articles. Compared to more formal genres, clinical language often does not comply with formal grammar. Moreover, misspellings abound and shorthand – abbreviations and acronyms – are often composed in an idiosyncratic and ad hoc manner, with some being difficult to disambiguate even in context [Liu *et al.*, 2001]. This may accidentally lead to an artificially high degree of synonymy and homonymy[1] – problematic phenomena from an information extraction perspective. In one study, it was found that the pharmaceutical product *noradrenaline* had approximately sixty spelling variants in a set of intensive care unit nursing narratives written in Swedish [Allvin *et al.*, 2011]. This example may serve to illustrate some of the challenges involved in processing clinical language.

There are many challenges in applying machine learning and natural language processing in the healthcare domain. One important challenge is that the data is potentially high-dimensional and inherently sparse. The high dimensionality is a consequence of the many variables that can be used to describe the examples, e.g., patients or healthcare episodes: the number of unique diagnosis codes and drug codes is, for instance, huge, and the problem is further exacerbated when incorporating text data in the predictive models. The ensuing sparsity can be

[1] Homonyms are two orthographically identical words (same spelling) with different meanings. An example of this phenomenon in Swedish clinical text is *pat*, which is short for both *patient* and *pathological*.

explained by the fact that only a very small subset of the clinical events – or of the words in the vocabulary, in the case of using clinical notes – will be present in a patient or healthcare episode. This phenomenon is often referred to as the curse of dimensionality and makes it problematic to learn high-performing predictive models. Another challenge concerns the heterogeneity of EHR data, which does not allow the distinct types of clinical data to be treated in an identical manner. It is moreover non-trivial to exploit and combine the various data types for the purpose of predictive modeling. As a result many efforts to apply machine learning and natural language processing in the healthcare domain have tended to focus on only one of these broad categories of data, effectively failing to capitalize on their complementary nature.

4.2 SYNONYM EXTRACTION

Natural language processing systems need to be able to account for language use variability, which is often realized through word choice and made possible by a linguistic phenomenon known as synonymy. Synonymy is a semantic relation between two phonologically distinct words with very similar meanings. However, as words typically do not have the exact same meanings, we typically speak of near synonyms instead, which means that two words are interchangeable in some, but not all, contexts [Saeed, 1997]. In the healthcare domain, there are differences in the vocabulary used by clinicians and patients, where the latter use layman variants of medical terms [Leroy and Chen, 2001]. When processing clinical notes, it is important to have ready access to terminological resources that cover this variation in the use of vocabulary. Some examples of applications where this is particularly important include query expansion [Leroy and Chen, 2001], text simplification [Leroy *et al.*, 2012] and information extraction [Eriksson *et al.*, 2013a].

The importance of synonym learning is well recognized in the NLP research community, especially in the biomedical [Cohen and Hersh, 2005] and clinical [Meystre *et al.*, 2008] domains. A whole host of techniques have been proposed for the identification of synonyms and other semantic relations, including the use of lexico-syntactic patterns, graph-based models and, indeed, distributional semantics [Henriksson *et al.*, 2012, 2013a,b; Skeppstedt *et al.*, 2013; Jagannatha *et al.*, 2015].

4.3 NAMED ENTITY RECOGNITION

Named entity recognition (NER) concerns the ability to recognize references to entities of certain predefined semantic categories in free-text. This ability is a key enabler of accurate information extraction and is a basic building-block of many, more sophisticated, NLP systems. The importance of NER in the healthcare domain is well recognized; testament to that are the many shared tasks and challenges that have been organized in recent years on automatic recognition of medical entities in clinical text [Uzuner *et al.*, 2010, 2011a; Pradhan *et al.*, 2014, 2015]. While most of the NER modules that are currently in use in clinical NLP systems are rule-based and rely heavily on the existence of comprehensive medical dictionaries, the trend is moving in the direction of machine learning, with which state-of-the-art results have been obtained [de Bruijn *et al.*, 2011; Tang *et al.*, 2013; Skeppstedt, 2014].

Named entity recognition is typically approached as a structured prediction task, i.e., when sequential data needs to be segmented and/or labeled. Conditional random fields (CRF) [Lafferty *et al.*, 2001] is a popular choice of learning algorithm for NER. The power of CRF lies in its ability to model multiple variables that are dependent on each other – as they typically are in structured prediction tasks – and to exploit large sets of input features. It achieves this by using an undirected probabilistic graphical model that, unlike, e.g., Hidden Markov Models, is discriminative. IOB-encoding is often used for encoding the output variables, which indicates whether a token is at the beginning (B), inside (I) or outside (O) a given named entity mention.

4.4 ADVERSE DRUG EVENT DETECTION

Adverse drug events (ADEs) constitute a major public health issue. In fact, ADEs have been estimated to be responsible for approximately 3-5% of hospital admissions worldwide [Beijer and De Blaey, 2002; Howard *et al.*, 2007]. This causes inflated healthcare costs and suffering in patients, often unnecessarily so as many ADEs are preventable [Hakkarainen *et al.*, 2012]. Pharmacovigilance is therefore carried out throughout the life-cycle of a drug in order to inform decisions on its initial and sustained use in the treatment of patients. The need

to monitor the safety of drugs post marketing is caused by the inherent limitations of clinical trails in terms of sample size and study duration, making it particularly difficult to identify rare and long-latency ADEs. There are, in fact, several cases in which drugs have been discovered to cause severe, even fatal, ADEs, resulting in their withdrawal from the market [Sibbald, 2004; Furberg and Pitt, 2001]. Post-marketing surveillance of drug safety has primarily relied on case reports that are reported voluntarily by clinicians and drug users in so-called spontaneous reporting systems, such as the US Food and Drug Administration's Adverse Event Reporting System, the Yellow Card Scheme in the UK and the World Health Organization's Global Individual Case Safety Reporting Database: Vigibase. Relying solely on spontaneous reports has, however, proven to be insufficient. In addition to several limitations inherent in collecting information in this way, such as selective reporting, incomplete patient information and indeterminate population information [Goldman, 1998], spontaneous reporting systems suffer heavily from underreporting: according to one estimate, more than 94% of ADEs are not reported in such systems [Hazell and Shakir, 2006].

As a result, alternative sources of information for pharmacovigilance are being explored, including the biomedical literature [Van Mulligen *et al.*, 2012], user-generated data in social media [Sarker *et al.*, 2015] and EHRs. The latter has the distinct advantage of containing data collected from the clinical setting, thereby providing access to longitudinal observations of patients, their medical condition and drug use. Detection of ADEs in EHRs has primarily focused on using either structured [Harpaz *et al.*, 2010; Chazard *et al.*, 2011; Karlsson *et al.*, 2013; Zhao *et al.*, 2014b,a, 2015a,b] or unstructured data [Eriksson *et al.*, 2013b; LePendu *et al.*, 2013].

PART II

METHODOLOGY

CHAPTER 5

METHODOLOGY

In this chapter, the research strategy will be described from several perspectives, including a brief discussion of the philosophical assumptions that underpin it. This will help to explain and motivate the general research strategy assumed in the thesis.

After a description of the research strategy, including a characterization of the research and a discussion around the underlying assumptions, the evaluation framework – based on, and an integral part of, the former – is described out in more detail. Finally, ethical issues that arise when conducting research on clinical data are discussed briefly. Much of this is based on [Henriksson, 2013].

5.1 RESEARCH STRATEGY

The research strategy (see Figure 5.1) assumed in this thesis is predominantly quantitative in nature, with the evaluations being conducted using statistical measures. The research can moreover be characterized as deductive: a hypothesis is formulated and verified empirically and quantitatively, followed by hypothesis testing. In hypothesis testing, conclusions are drawn using statistical (frequentist) inference. The hypothesis we wish to test is called the null hypothesis and is rejected or accepted according to some level of significance. When comparing two classifiers, statistical significance tests essentially determine, with a given degree of confidence, whether the observed difference in performance (according to some metric) is significant [Alpaydin, 2010].

The research can be characterized as exploratory on one level and causal on another. Strategies for creating semantic space ensembles are investigated in an exploratory fashion; there are certainly other strategies that could perhaps equally well have been investigated. The research is also, to a large extent, causal: in many of the included studies we are investigating the effect of one variable – the introduction of some technique or method – on another. The techniques or methods are evaluated in an experimental setting. In fact, the experimental approach underlies much of the general research strategy. This is a well-established means of evaluating NLP systems and adheres to the Cranfield paradigm, which has its roots in experiments that were carried out in the 1960s to evaluate indexing systems [Voorhees, 2002]. In this evaluation paradigm, two or more systems are evaluated on identical data sets using statistical measures of performance. There are two broad categories of evaluation: system evaluation and user-based evaluation. In contrast to user-based evaluations, which involve users and more directly measure a pre-specified target group's satisfaction with a system, system evaluations circumvent the need for direct access to large user groups by creating reference standards or gold standards. These are pre-created data sets that are treated as the ground truth, to which the systems are then compared.

In quantitative research, a positivist philosophical standpoint is often assumed, to which an interpretivist[1] position is then juxtaposed. While these positions are usefully opposed[2], many do not subscribe wholly and unreservedly to either of

[1] Interpretivism is sometimes referred to as antipositivism.

[2] Some have, however, argued that these can be combined in a single study [Myers, 1997].

the views. In the Cranfield paradigm, with its experimental approach, there is certainly a predominance of positivist elements. This is naturally also the case when performing predictive modeling. As predictive models are necessarily based on a logic of hypothetico-deduction, research that involves predictive modeling is also based on positivist assumptions.

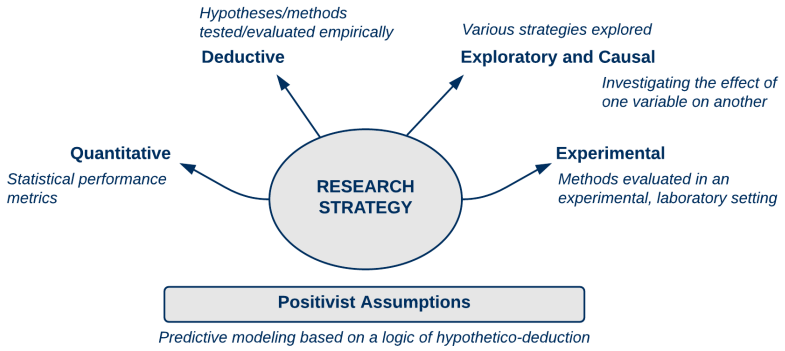

Figure 5.1: RESEARCH STRATEGY; ADOPTED FROM [HENRIKSSON, 2013]

5.2 EVALUATION FRAMEWORK

As mentioned earlier, the evaluation is mainly conducted quantitatively with statistical measures of performance in an experimental setting, where one system is compared to another, which is sometimes referred to as the baseline (system). The hypothesis is often in the simple form of:

```
System A will perform better than System B on Task T given Performance Metric M,
```

even if it is not always explicitly stated as such. The performance measures are commonly calculated by comparing the output of a system with a reference standard, which represents the desired output as defined by human experts. The automatic evaluations conducted in the included studies are extrinsic – as opposed to intrinsic – which means that the semantic spaces, along with their various ways

of representing meaning, are evaluated according to their utility, more specifically their ability to perform a predefined task [Resnik and Lin, 2013].

5.3 PERFORMANCE METRICS

Widely used statistical measures of performance in information retrieval and NLP are precision and recall [Alpaydin, 2010]. To calculate these, the number of true positives (tp), false positives (fp) and false negatives (fn) are needed. For some measures, the number of true negatives (tn) is also needed. True positives are correctly labeled instances, false positives are instances that have been incorrectly labeled as positive and false negatives are instances that have been incorrectly labeled as negative. Precision (P), or positive predictive value (PPV), is the proportion of positively labeled instances that are correct:

$$Precision : P = \frac{tp}{tp + fp} \tag{5.1}$$

Recall (R), or *sensitivity*, is the proportion of all actually positive instances that are correctly labeled as such:

$$Recall : R = \frac{tp}{tp + fn} \tag{5.2}$$

There is generally a tradeoff between precision and recall: by assigning more instances to a given class (i.e., more positives), recall may, as a consequence, increase – some of these may be true (tp) – while precision may decrease since it is likely that the number of false positives will, simultaneously, increase. Depending on the application and one's priorities, one can choose to optimize a system for either precision or recall. F-score incorporates both notions and can be weighted to prioritize either of the measures:

$$F\text{-}score : F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R} \tag{5.3}$$

When one does not have a particular preference, F_1-score is often employed and is defined as the harmonic mean of precision and recall.

In Paper I, the semantic space ensembles are used for k nearest neighbors retrieval. Here, recall among the k nearest neighbors is calculated as follows:

$$\textit{Recall (top k)} : R_{\textit{top k}} = \frac{tp}{tp+fn} \text{in a list of } k \text{ labeled instances} \tag{5.4}$$

A form of weighted precision is calculated as follows:

$$\begin{aligned} \textit{Weighted Precision} : P_w &= \frac{\sum_{i=0}^{k-1} (k-i) \cdot f(i)}{\sum_{i=0}^{k-1} j-i} \\ &\text{where} \\ f(i) &= \begin{cases} 1 & \text{if } i \in \{tp\} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5.5}$$

In words, this assigns a score to true positives according to their (reverse) ranking in the list, sums their scores and divides the total score by the maximum possible score (where all k labels are true positives).

In several of the studies, accuracy and area under the ROC curve (AUC) are considered. Accuracy corresponds to the percentage of correctly classified instances:

$$\textit{Accuracy} : A = \frac{tp+tn}{tp+fp+fn+tn} \tag{5.6}$$

AUC depicts the performance of a model without regard to class distribution or error costs by estimating the probability that a model ranks a randomly chosen positive instance ahead of a negative one.

5.4 RELIABILITY AND GENERALIZATION

Metrics like the ones described above are useful in estimating the expected performance – measured according to one or more criteria – of a system. There are, however, two closely related aspects of system evaluations that are important to consider: (1) reliability, i.e., the extent to which we can rely on the observed performance of a system, and (2) generalization, i.e., the extent to which we can expect the observed performance to be maintained when we apply the system to unseen data.

The reliability of a performance estimation depends fundamentally on the size of the data used in the evaluation: the more instances we have in our (presumably representative) evaluation data, the greater our statistical evidence is and thus the more confident we can be about our results. This can be estimated by calculating a confidence interval or, when comparing two systems, by conducting a statistical significance test.

However, in order to estimate the performance of a system on unseen data – as we wish to do whenever prediction is the aim – we need to ensure that our evaluation data is distinct from the data we use for training our model. The goal, then, is to fit our model to the training data while eschewing both underfitting – where the model is less complex than the function underlying the data – and overfitting, where the model is too complex and fitted to peculiarities of the training data rather than the underlying function. In order to estimate generalization ability and, ultimately, expected performance on unseen data, one can, for instance, partition the data into three: a training set, a development set and an evaluation set. The training set, as the name suggests, is used for training the model; the generalization ability of a model typically improves with the number of training instances [Alpaydin, 2010]. The development set is used for model selection and for estimating generalization ability: the model with the highest performance (and the one with the best inductive bias) on the development set is the best one. This is done in a process known as cross-validation. In order to estimate performance on unseen data, however, the best model needs to be re-evaluated on the evaluation set, as the development set has effectively become part of the training set[3] by using it for model selection. When conducting comparative

[3]Sometimes the training and development sets are jointly referred to as the training set and the evaluation set as the test set.

evaluations, all modifications to the (use of the) learning algorithm, including parameter tuning, must be made prior to seeing the evaluation data. Failing to do so may lead to the reporting of results that are overly optimistic [Salzberg, 1997]. An alternative is to perform *k*-fold cross-validation, where a common choice for *k* is 10. In 10-fold cross-validation, the dataset is partitioned into ten equally-sized parts: nine parts are used for training and one part is used for testing; this process is repeated ten times, each time using a different part for testing. To obtain more reliable performance estimates, *k*-fold cross-validation is often repeated multiple times (e.g., 10), each time with a different partitioning into folds.

It is also worth underscoring the fact that empirical evaluations of this kind are in no way domain independent: any conclusions we draw from our analyses – based on our empirical results – are necessarily conditioned on the data set we use and the task we apply our models to. The *No Free Lunch Theorem* [Wolpert, 1995] states that there is no such thing as an optimal learning algorithm – for some data sets it will be accurate and for others it will be poor. By the same token, it seems unlikely to believe that there is a single optimal way to model semantics; the results reported here are dependent on a specific domain (or data set) and particular applications.

5.5 STOCKHOLM EPR CORPUS

All of the included studies uses data extracted from the Stockholm EPR Corpus [Dalianis *et al.*, 2009, 2012], which is a database of electronic health records[4] from Karolinska University Hospital. The Stockholm EPR Corpus contains the health records of around 700,000 patients in the Stockholm region over several years; in most of the studies, data was extracted from the years 2009 and 2010. More information can be found in the papers.

[4]This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

5.6 ETHICAL ISSUES

Working with clinical data inevitably raises certain ethical issues. This is due to the inherently sensitive nature of this type of data, which is protected by law. In Sweden, there are primarily three laws that, in one way or another, govern and stipulate the lawful handling of patient information. The Act Concerning the Ethical Review of Research Involving Humans[5], SFS 2003:460 [Sveriges riksdag, 2013a], is applicable also to research that involves sensitive personal data (3 §). It states that such research is only permissible after approval from an ethical review board (6 §) and that processing of sensitive personal data may only be approved if it is deemed necessary to carry out the research (10 §). More specific to clinical data is the Patient Data Act[6], SFS 2008:355 [Sveriges riksdag, 2013b]. It stipulates that patient care must be documented (Chapter 3, 1 §) and that health records are an information source also for research (Chapter 3, 2 §). A related, but more generally applicable, law is the Personal Data Act[7], SFS 1998:204 [Sveriges riksdag, 2013c], which protects individuals against the violation of their personal integrity by processing of personal data (1 §). This law stipulates that personal data must not be processed for any purpose that is incompatible with that for which the information was collected; however, it moreover states that processing of personal data for historical, statistical or scientific purposes shall not be regarded as incompatible (9 §). If the processing of personal data has been approved by a research ethics committee, the stipulated preconditions shall be considered to have been met (19 §).

The Stockholm EPR Corpus does not contain any structured information that may be used directly to identify a patient. The health records were in this sense de-identified by the hospital by replacing social security numbers (in Swedish: personnummer) with a random key (one per unique patient) and by removing all names in the structured part of the records. The data is encrypted and stored securely on machines that are not connected to any networks and in a location to which only a limited number of people, having first signed a confidentiality agreement, have access. Published results contain no confidential or other information that can be used to identify individuals. The research conducted in this thesis has been approved by the Regional Ethical Review

[5] In Swedish: Lag om etikprövning av forskning som avser människor
[6] In Swedish: Patientdatalagen
[7] In Swedish: Personuppgiftslagen

Board in Stockholm[8] (Etikprövningsnämnden i Stockholm), permission numbers 2009/1742-31/5 and 2012/834-31/5. When applying for these permissions, the aforementioned measures to ensure that personal integrity would be respected were carefully described, along with clearly stated research goals and descriptions of how, and for what purposes, the data would be used.

Some of these ethical issues also impact on scientific aspects of the work. The sensitive nature of the data makes it difficult to make data sets publicly available, which, in turn, makes the studies less readily *repeatable* for external researchers. Developed techniques and methods can, however, be applied and compared on clinical data sets to which a researcher does have access. Unfortunately, access to large repositories of clinical data for research purposes is still rather limited, which also means that many problems have not yet been tackled in this particular domain and – by extension – that there are fewer methods to which your proposed method can be compared. This is partly reflected in this work, where an external baseline system is sometimes lacking. Without a reference (baseline) system, statistical inference – in the form of testing the statistical significance of the difference between the performances of two systems – cannot be used to draw conclusions. Employing a more naïve baseline (such as a system based on randomized or majority class classifications) for the mere sake of conducting significance testing is arguably not very meaningful.

[8] http://www.epn.se/en/

PART III

EMPIRICAL INVESTIGATIONS

CHAPTER 6

SYNONYM EXTRACTION

In this chapter, the notion of semantic space ensembles is explored in the context of a natural language processing task for which the use of distributional semantics comes naturally: the task of assessing the semantic similarity between pairs of words and, more specifically, extracting candidate synonyms for a given term. Here, the – in some sense – related task of identifying the abbreviated versus expanded form a word. Ensembles of semantic spaces are created by utilizing both model hyperparameters and different types of corpora.

6.1 BACKGROUND

Models of distributional semantics are, in some sense, able to capture the semantics underlying word choice. This quality of natural language is essential and endows it with the flexibility we, as humans, need in order to express and communicate our thoughts. For computers, however, language use variability is problematic and far removed from the unambiguous, artificial languages to which they are used. There is thus a need for computers to be able to handle and account for language use variability, not least by being aware of the relationship between surface tokens and their underlying meaning, as well as the relationship between the meaning of tokens. Indeed, morphological variants, abbreviations, acronyms, misspellings and synonyms – although different in form – may share semantic content to different degrees. The various lexical instantiations of a concept thus need to be mapped to some standard representation of the concept, either by converting the different expressions to a canonical form or by generating lexical variants of a concept's preferred term. These mappings are typically encoded in semantic resources, such as thesauri or ontologies, which enable the recall of information extraction systems to be improved. Although their value is undisputed, manual construction of such resources is often prohibitively expensive and may also result in limited coverage, particularly in the biomedical and clinical domains where language use variability is exceptionally high [Meystre *et al.*, 2008; Allvin *et al.*, 2011].

6.2 SINGLE AND MULTIPLE-CORPORA ENSEMBLES

In Paper I, two strategies for creating semantic space ensembles are explored:

1. Combining semantic spaces that have been constructed with different strategies for handling word order within the context window
2. Combining semantic spaces that have been constructed from different types of corpora

In addition to the word order strategy, another hyperparameter is varied, namely the size of the context window. In this study, however, this hyperparameter is

not employed as a strategy for creating semantic space ensembles; instead, it is optimized in the conventional manner – both for individual semantic spaces and for semantic space ensembles – in the search for high-performing models. Single-corpus and multiple-corpora ensembles are compared to each other, as well as to individual semantic spaces. The single-corpus ensembles are composed of two semantic spaces, wherein each employs a different word order strategy, while the multiple-corpora ensembles are composed of up to four semantic spaces, where each semantic space is constructed with a unique combination of word order strategy and underlying corpus. In terms of corpus preprocessing, when constructing semantic spaces that take into account word order, the models are moreover optimized for whether stop words should be retained or removed. Stop-word filtering is sometimes applied prior to semantic space construction since co-occurrence information involving high-frequent and widely distributed words contribute very little to determining the meaning of words. However, when taking into account word order, it may be counterproductive to remove stop words, as, for instance, function words – which are typical members of stop-word lists – are important to the syntactic structure of language and may thus be of value when modeling order information. Once the semantic spaces that are to be included in an ensemble have been selected, the question of how to combine them arises – various strategies for combining the constituent semantic spaces were explored.

6.3 MAIN FINDINGS

A combination of two distributional models – Random Indexing and Random Permutation – employed in conjunction with a single corpus outperforms using either of the models in isolation. Furthermore, combining semantic spaces induced from different types of corpora – a corpus of clinical text and a corpus of medical journal articles – further improves results, outperforming a combination of semantic spaces induced from a single source, as well as a single semantic space induced from the conjoint corpus. A combination strategy that simply sums the cosine similarity scores of candidate terms is generally the most profitable out of the ones explored.

This study demonstrates that ensembles of semantic spaces can yield improved performance on the tasks of automatically extracting synonyms and abbreviation-expansion pairs. This notion, which merits further exploration,

allows different distributional models – with different model parameters – and different types of corpora to be combined, potentially allowing enhanced performance to be obtained on a wide range of natural language processing tasks.

CHAPTER 7

NAMED ENTITY RECOGNITION

In this chapter, ensembles of semantic spaces are explored in the context of named entity recognition. The proposed method is based on a form of learning known as *prototype learning*, wherein canonical examples of each class are provided. In Paper II, III and VI, a semi-supervised approach is taken: a corpus annotated for named entities is assumed to be available, in conjunction with a large unannotated corpus. The available annotations are exploited to learn prototypical representations of each named entity class in semantic space; ensembles of distributed prototypes are created by tinkering with model hyperparameters: different strategies for handling word order and different context window sizes.

7.1 BACKGROUND

Named entity recognition (NER) concerns the ability to recognize references to entities of certain predefined semantic categories in free-text. This ability is a key enabler of accurate information extraction, which has grown in importance with the inexorably growing amounts of digitized data. One application area for information extraction that is receiving considerable attention at the moment is healthcare, where great amounts of data are now being stored as a result of the increasing adoption of electronic health records (EHRs). Since the majority of this data is in the form of text, information extraction and other natural language processing (NLP) methods need to be adapted to this particular domain. This is especially challenging due to the properties of clinical text: formal grammar is typically not complied with, while misspellings and non-standard shorthand abound [Allvin *et al.*, 2011]. Testament to the growing importance of domain-adapted NER systems are the many shared tasks and challenges that have been organized in recent years [Uzuner *et al.*, 2010, 2011b; Pradhan *et al.*, 2014, 2015]. However, most of the existing NER modules that are used in clinical NLP systems, such as MedLEE [Friedman, 1997], MetaMap [Aronson and Lang, 2010] and cTAKES [Savova *et al.*, 2010], are rule-based – i.e., with hand-crafted rules – and thus rely heavily on comprehensive medical dictionaries. The trend is, however, increasingly moving in the direction of machine learning, with state-of-the-art clinical NER systems being primarily based on predictive models [de Bruijn *et al.*, 2011; Tang *et al.*, 2013; Zhang *et al.*, 2014].

7.2 ENSEMBLES OF DISTRIBUTED PROTOTYPES

The proposed method presupposes the availability of two resources: (1) an annotated (named entity) corpus and (2) an unannotated corpus. While the annotated corpus may be relatively small, the unannotated corpus should preferably be much larger and in the same domain. The method essentially consists of the following steps:

1. Learning multiple distributed prototypes for each semantic category

2. Generating features for the instances (words) based on their distance in semantic space to each of the prototype vectors
3. Applying an appropriate learning algorithm to the annotated corpus with a feature set that includes the generated features

The core of the method is in the first two steps, which concern the provision of semantic features to the learning algorithm with the use of distributed word representations.

LEARNING PROTOTYPES

A distributed prototype vector is an abstract representation of a target (named entity) class. It is learned by exploiting the existing annotations to obtain their (distributed) representations in semantic space, which is constructed over a large, unannotated corpus. The prototype vector of a semantic category is then obtained by taking the centroid of the semantic vectors representing the category's annotated instances that occur above some threshold t in the unannotated corpus; here, t is set to a fairly large number: 100. Low-frequency terms are not included since the statistical foundation for their representation is weak, i.e., the observations of word usage are few. The centroid is defined as the median value of each dimension, as it was shown to lead to a better separation of classes compared to using the column-wise mean values. This results in an abstract representation, i.e., one that does not correspond to an actual instance, which is otherwise often the case when calculating the centroid of a cluster. When employing a set of semantic spaces, a prototype vector is obtained for each semantic space and semantic category (Algorithm 1).

Multiple semantic spaces are created by employing different word order strategies and different context window sizes.

GENERATING DISTRIBUTIONAL FEATURES

The prototype vectors are then used for generating features that describe the instances – which are here words or tokens – in the dataset. There is one binary feature per named entity class and semantic space, where the value is either True or

Algorithm 1: Learning Multiple Prototype Vectors for a Semantic Category

input : multiset W of seed words, set S of semantic spaces
output: set of n-dimensional distributed prototype vectors $P = \{\overrightarrow{p_1}, ..., \overrightarrow{p_{|S|}}\}$

```
for s ∈ S do
    for w ∈ W do
        v⃗ ← SemanticVector(w, s)
        /* append coordinate at position i in v⃗ to c_i          */
        for i ← 1 to n do
            Append(v⃗_i, c_i)
        end
    end
    /* get column-wise median values                            */
    for i ← 1 to n do
        p⃗_i ← Median(c_i)
    end
    /* append prototype vector from s to P                      */
    Append(p⃗, P)
end
return P
```

False depending on whether the cosine similarity between the target word and the prototype vector is above a set threshold. The threshold is based on the pairwise distances between the annotated named entities of a certain semantic category and its corresponding prototype vector in a given semantic space. The threshold is set to maximize F_β-score on the training set, where the positive examples are the instances that belong to a certain semantic category and the negative examples are all other instances (Equation 7.1).

$$\arg\max_{t \in \mathbb{V}} \left((1+\beta^2) \frac{P^{(t)} \cdot R^{(t)}}{(\beta^2 \cdot P^{(t)}) + R^{(t)}} \right), \tag{7.1}$$

where P is precision (true positives / true positives + false positives) and R is recall (true positives / true positives + false negatives); $\mathbb{V} = (0, 0.0001, 0.0002, ..., 1)$; β determines the weight that should be given to recall relative to precision. The lowest threshold is chosen that optimizes the F_β-score.

SEQUENCE LABELING

In addition to the generated distributional semantic features, a set of orthographic and syntactic features are also generated. These features, in addition to the generated semantic features, are then provided to the learning algorithm together with the class labels. Following the standard approach to training a NER model, we cast the problem as a sequence labeling task, in which the goal is to find the best sequence of labels for a given input, i.e., the sequence of tokens in a sentence, which are described by various features. IOB-encoding of class labels is used, which indicates whether a token is at the beginning (B), inside (I) or outside (O) a given named entity mention. The features, along with the class labels, are then provided to a learning algorithm.

7.3 MAIN FINDINGS

In Paper II, it was shown that generating features with prototype vectors lead to improved performance over employing a set of orthographic and syntactic features. It was also shown that further improvements could be gained by employing an ensemble of three semantic spaces, each employing a different word order strategy. Four different strategies for combining the semantic spaces were explored, out of which the most profitable one turned out being simply to fuse the features generated by each semantic space.

In Paper III, larger ensembles of semantic spaces were created, wherein not only the word order strategy was exploited but also the size of the context window. Altering both of these model hyperparameters led to the creation of more effective semantic space ensembles compared to only altering the word order strategy. Follow-up analyzes were also conducted, which revealed the contribution made by each respective window size and word order strategy.

CHAPTER 8

ADVERSE DRUG EVENT DETECTION

In this chapter, ensembles of semantic spaces are explored in the context of adverse drug event detection. There are two parts: in Paper IV and V, heterogeneous types of clinical data are modeled in semantic space to create representations of healthcare episodes; in Paper VI, learning to detect relations between drugs and symptoms/disorders is achieved with distributional semantic representations of context. The semantic space ensembles are created by using multiple context window sizes.

8.1 BACKGROUND

Adverse drug events (ADEs), defined as undesired harms resulting from the use of a drug [Nebeker *et al.*, 2004], are considered to be a major public health issue: they are the most common type of iatrogenic injury [Kohn *et al.*, 2000] and cause around 3.7% of hospital admissions worldwide [Howard *et al.*, 2007]. In Sweden, ADEs have been identified as the seventh most common cause of death [Wester *et al.*, 2008]. Although the safety of a drug is evaluated in clinical trials before it is made available on the market, post-marketing drug safety detection is still necessary, since clinical trials are often limited in terms of duration and number of patients involved. As a result, not all severe ADEs caused by a drug can be discovered prior to its launch. There are many cases of drugs being withdrawn from the market, such as Vioxx for its doubled risk of causing myocardial infarction [Sibbald, 2004] and Cerivastatin for causing fatal rhabdomyolysis [Furberg and Pitt, 2001]. Therefore, post-marketing detection of potential ADEs is of high necessity and importance.

Pharmacovigilance, also known as post-marketing drug safety surveillance, has conventionally relied on collecting information reported voluntarily by clinicians and users of the target drugs in order to inform decisions about their continued use in the treatment of patients. This individual case reporting, however, comes with many limitations, such as under-reporting, dubious reliability and the absence of the drug users' medical history [Goldman, 1998]. This has led to the exploration of alternative, and complementary, sources of information for pharmacovigilance, including user-generated data in social media [Sarker *et al.*, 2015] and EHRs. The latter has the distinct advantage of containing data collected from the real-word setting of clinical practice [Coloma *et al.*, 2013], effectively giving us access to longitudinal observations of patients, their medical condition and drug use. Adverse drug events are encoded in EHRs with a limited number of diagnosis codes; however, EHRs also suffer from under-reporting of ADEs [Hazell and Shakir, 2006]. This makes it difficult to obtain accurate incidence rates of ADEs, which, in turn, may harm patient safety. There is thus a need for systems that can analyze EHR data and alert clinicians or expert coders of potential ADEs and suggest the corresponding diagnosis code to assign. It is also valuable to be able to detection mentions of ADEs in clinical notes; this requires relations to be detected and accurately labeled between mentions of drugs and disorders/symptoms.

8.2 ENSEMBLES OF HETEROGENEOUS DATA

To create representation of care episodes, we first generalize the framework of distributional semantics – traditionally employed with natural language data – to other types of sequential data. As most uses of distributional semantic models employ a context window around the target word, the data needs to be presented as a sequence. For each of the three structured data types, we extract all sequences of events that occur in the care episode of a patient, ordered by time. These sequences are then processed one-by-one by the distributional semantics algorithm. As structured data events are ordered by time, we have the possibility of creating interleaved sequences of these events. Notes cannot, however, readily be interleaved with the structure data events, as they themselves consist of sequences of words. The preprocessed notes are processed sentence-by-sentence. The semantic spaces can then be used to create feature representations of care episodes that are fed to the learning algorithm. This is achieved by adding up the semantic vectors of the items (words and events) in a given care episode.

8.3 ENSEMBLES OF CONTEXT REPRESENTATIONS

For learning to detect and label relations between named entities – primarily between drugs and disorders/symptoms – we study different representations of context to the left, right and in between pairs of entities. Three representations are compared: (1) as a bag of words, (2) as a bag of semantic vectors from a single semantic space, and (3) as bags of semantic vectors from multiple semantic spaces, where each semantic space has been constructed with a different context window size.

8.4 MAIN FINDINGS

In Paper IV, wherein the task of detecting ADEs in healthcare episodes was tackled, using the proposed bag-of-concepts representation yielded significant improvements in predictive performance over a bag-of-items representation for diagnosis codes and when combining structured and unstructured data. It was

also shown that combining structured and unstructured data led to significant improvements on the task of detecting adverse drug events in electronic health records, compared to using only one of the data types. In Paper V, it was demonstrated that improved predictive performance can be obtained on the task of detecting ADEs by creating representations of care episodes with the use of multiple distributional semantic spaces, each built with a different context window size. It was also shown that an early (feature) fusion was more successful than various late (classifier) fusion approaches.

PART IV

EPILOGUE

CHAPTER 9

DISCUSSION

In this chapter, we take a step back and embark on a wider discussion of the investigated strategies for creating ensembles of semantic spaces – both in terms of creating potentially diverse semantic spaces and in terms of combining the semantic spaces – that are effective, in the sense of outperforming the use of any of the individual semantic spaces. The usefulness of semantic space ensembles in general and the proposed methods in particular is then discussed. Finally, the use of semantic space ensembles in the investigated healthcare applications is summarized and it is argued that they have added utility in this particular domain.

9.1 CREATING SEMANTIC SPACES

In order to create effective ensembles, it is fundamental to create accurate and diverse base models [Dietterich, 2000a]. In this dissertation, two broad types of strategies were investigated in order to create distinct and potentially diverse – semantic spaces: (1) changing the underlying data, and (2) changing model hyperparameter configurations.

Changing the underlying data was done in two very different ways: (1) creating multiple representations of words from different sets of language use observations (i.e., different corpora), and (2) modeling different types of data. In the first case, two corpora from the same general domain – medicine – but from different genres – clinical text and journal articles – were utilized. This allowed a large variety of observational data on language use, where the contexts may be very different, to be exploited. This approach is particularly valuable when large in-domain corpora are not readily available; out-domain corpora can then be leveraged. In this study, corpora from similar domains but different genres were combined; it would, however, also be interesting to combine corpora from very different domains. In the second case, different types of data were modeled in semantic space. This made it possible to exploit various sources of information that in many ways are complementary. Although this can also be achieved without the use of distributional semantics – by, for instance, representing an example as a bag of items – the data would then potentially have a much higher dimensionality and be much more sparse; as we know, learning high-performing predictive models from such data is known to be challenging.

Two model hyperparameters were in this dissertation altered to create distinct semantic spaces: (1) the strategy for handling word order in random indexing, and (2) the size of the context window. The different strategies for handling word order each have their pros and cons. By employing a strict definition of co-occurrence events, wherein a co-occurrence of a certain type is defined according to its exact position in relation to the target word, word order is given a bigger role in deriving word meaning; however, it fails to account for any similarity whatsoever between co-occurrence events in different positions. This results in denser semantic vectors and also requires that the dimensionality is sufficiently large to ensure that the near-orthogonality property of random indexing is maintained. By employing a bag-of-words strategy within the context window, word order is effectively ignored. This results in sparser semantic vectors and does

not require the dimensionality to be as high as when employing order vectors. Direction vectors provide a middle ground by taking into account the direction in relation to the target word, i.e., whether the word appears to the left or right of the target word. The overlap between the nearest neighbors of words in these different spaces has previously been shown to be small [Sahlgren *et al.*, 2008]. This indicates that, to some extent, diverse semantic spaces can be created, which, given that the different semantic spaces are "accurate", allows them to be combined to yield effective ensembles. This hyperparameter was manipulated to create semantic space ensembles in Paper I, II and III. The diversity of semantic spaces created with different strategies for handling word order was evaluated in several different ways in Paper III. Not only was it shown that combining semantic spaces constructed with different word order strategies led to improved predictive performance on named entity recognition, but also that the local neighborhoods of prototype vectors inhabiting the different spaces are rather different, producing largely independent rankings of top-1000 nearest neighbors.

Changing the size of the context window also affects the composition of the resulting semantic space. This was exploited to create effective semantic space ensembles in Paper IV, V and VI. On the one hand, this allows different semantic relations to be captured. On the other hand, it also circumvents the problem of finding a single optimal window size. Even if such a window size exists *overall*, it is unlikely to be optimal for each and every instance in a dataset. This approach does not, however, remove the window size hyperparameter, and the question remains as to which window sizes should be employed in order to create the most effective semantic space ensemble. In Paper V it was shown that adding more semantic spaces to the ensemble generally leads to improved predictive performance; however, the improvement was not monotonic and did not appear particularly stable. This is in contrast to other ensemble models like random forest, where, as a rule of thumb, the more base models that are included in the ensemble, the better. Whether the same behavior would be observed when employing semantic space ensembles directly for k nearest neighbor retrieval remains to be investigated.

There are, of course, other model hyperparameters that could be manipulated to create semantic spaces for inclusion in an ensemble, as well as additional ways of manipulating the underlying data. This represents a direction of research that could be taken in future work (see more in section 10.3).

9.2 COMBINING SEMANTIC SPACES

Once a distinct and potentially diverse set of semantic spaces has been generated, the subsequent step is to combine them in some way. Although this is not the focus of the work presented in this thesis, the question of how to combine the semantic spaces is unavoidable. Here, the semantic space ensembles were used in various ways: (1) directly, for k nearest neighbor retrieval, and (2) indirectly, for generating features that were exploited by a supervised learning algorithm. The way in which the ensembles are used influences what the appropriate combination strategies may be.

In the former case, in Paper I, strategies were considered that involved combining ranked lists. The items were ranked according to cosine similarity scores, which means that there are two types of information to exploit: (1) the ranking and (2) cosine similarity scores. The rankings produced by the various semantic spaces do not necessarily have to treated uniformly; in fact, several strategies treated the rankings produced by semantic spaces using different word order strategies asymmetrically. The most successful combination strategy was, however, simply to add up the cosine similarity scores in each of the semantic spaces and return a re-ranked list based on the summed values. It makes sense that cosine similarity is valuable information that should be exploited when combining multiple ranked lists. There are of course many other strategies that could be evaluated in future work, for instance taking the max (or min) observed cosine similarity score in the semantic spaces, as well as many more sophisticated methods. One option could be to learn from data how best to combine the output of the constituent semantic spaces. However, as described in the background on ensemble models, an advantage of unsupervised combination strategies is that they do not require labeled data and are not susceptible prone to overfitting.

In the latter case, various early (feature) and late (classifier) fusion strategies were considered. In Paper II, four feature fusion strategies were compared for named entity recognition using conditional random fields for generating predictive models. The strategy that led to the best predictive performance involved simply concatenating features generated by each semantic space in the ensemble. Compared to the other strategies that were considered, this strategy results in larger feature space. Given the experimental results, one, however, has to conclude that the learning algorithm was able to exploit the additional features to yield higher predictive performance. In Paper III, further analyses

were conducted that revealed, not only that certain types of semantic spaces – constructed with different hyperparameter configurations – led to better features than others, but also that removing a category of semantic spaces did not improve (micro-averaged) predictive performance. This consolidates the finding from Paper II that providing features from larger semantic space ensembles seems to lead to better predictive performance compared to using smaller ensembles. In Paper V early fusion, in the most basic sense of concatenating features from all semantic spaces in the ensemble, was compared to various late fusion strategies, where separate random forest classifiers were trained using features generated with semantic spaces constructed with different context window sizes. The experimental results were unequivocal, with the simple feature fusion approach outperformed all considered classifier fusion strategies. Here, the classifier fusion strategies involved combining the output of multiple random forest classifiers; in future work, it would interesting if the conclusion would be reached with other learning algorithms. A possible explanation for the relative success of the feature fusion approach is that it allows variable interactions to be exploited: combinations of features generated by the various semantic spaces in the ensemble may interact. Here, feature fusion was approached in the most straightforward manner; in future work, it would be interesting to explore other ways of fusing features generated by different semantic spaces; see, for instance, [Yang *et al.*, 2003].

9.3 SEMANTIC SPACE ENSEMBLES

The notion of semantic space ensembles proved to be promising in the empirical investigations in terms of improving the predictive performance on the tasks in which they were used. This is encouraging for several reasons, not least in that it allows unsupervised methods to be better capitalized on. This is often of great value, particularly in low-resource settings where access to large labeled datasets is scarce. As large amounts of unlabeled tend to be readily available in most domains, it is substantially more cost-effective to use slightly more computationally expensive unsupervised methods – caused by using an ensemble as opposed to a single semantic space – than manually to annotate more data.

To some degree, semantic space ensembles also reduce the need to optimize model hyperparameters carefully, which is cumbersome and requires additional held-out tuning data. It is clear that properly configuring the hyperparameters

of distributional semantic models is important and tends to result in significant differences when employing the obtained word representations in downstream tasks. By using an ensemble of semantic spaces, where the individual semantic spaces have been trained with different hyperparameter configurations, the potentially negative impact of choosing suboptimal hyperparameters is effectively reduced. The question of whether subsets of semantic spaces can be identified, following the overproduce-and-select paradigm of ensemble learning [Zhou *et al.*, 2002], remains to be investigated.

Modeling data in semantic space has many distinct advantages. On the one hand, it mitigates the problem of learning from high-dimensional and sparse data by creating dense, reduced-dimensional representations. Even when using ensembles of semantic spaces in combination with feature fusion, the dimensionality of the feature space is dependent only on the number of semantic spaces and their dimensionality. In that sense, it is inherently scalable, as the dimensionality does not grow with the size of the data, which entails that more information can readily be incorporated into the learning process. In Paper IV, V and VI, the semantic vectors of the corresponding items in each example were simply summed; in future work, alternatives to this could be explored. In Paper IV and V, the framework of distributional semantics was extended to various structured data, which provided a straightforward means of combining heterogeneous data without ending up with exceedingly high-dimensional feature space. Modeling data in semantic space also has the distinct advantage of taking into account and explicitly modeling similarities between variables, not only words but also diagnosis codes, drug codes and types of clinical measurements. This also contributes to alleviating potential data sparsity issues in the labeled dataset, as similarities are modeled in an entirely unsupervised setting. In Paper II and III, however, prototypical representations of semantic categories were learned in semantic space by also leveraging a small amount of labeled examples.

9.4 APPLICATIONS IN HEALTHCARE

Although the notion of semantic space ensembles is generalizable across domains and languages, in this thesis they were employed in tasks with promising applications in healthcare. In this domain, it is particularly important to be able to capitalize fully on unsupervised methods, as the costs involved in creating

annotated resources for supervised machine learning is higher than in many other domains. In Paper I, ensembles of semantic spaces were employed in an unsupervised setting; in Paper II, III, IV, V, VI, ensembles of semantic spaces were evaluated in a semi-supervised setting. Ensembles of semantic spaces moreover allow several problems, which tend to be present when working with data from electronic health records, to be mitigated: learning from heterogeneous, high-dimensional and sparse data. These challenges and how semantic space ensembles address them have been described in section 9.3.

Synonym extraction, addressed in Paper I, is important in the domains of biomedicine and healthcare in order to facilitate the creation of domain ontologies and terminologies. The large degree of language use variation in clinical text makes it paramount, for the second of meaningful secondary use of electronic health records, to have access to wide-coverage synonym and abbreviation lists. Distributional semantics can naturally be used for this purpose, as both synonyms and abbreviation-word pairs are prime examples of the paradigmatic relations captured by distributional semantic models. By using semantic space ensembles instead of a single semantic space, the predictive performance on this tasks was significantly improved. In this context, the semantic space ensembles were created not only by manipulating the model hyperparameters but also by exploiting an out-domain corpus. The encouraging results motivate future studies to investigate whether additional improvements can be gained by, for instance, exploiting corpora from additional domains or by creating more domain-specific (sub-)corpora. It would be interesting to create sub-corpora based on clinical practice, for instance, so that the various nuances of words, based on their different professional uses, can be better captured.

Named entity recognition, addressed in Paper II, III and VI, is a basic building-block of many natural language processing systems; it likewise represents an important task in the domain of healthcare and is a prerequisite for many more complex tasks such as relation extraction. The semi-supervised method developed in this thesis requires access to a small amount of labeled examples for each target named entity class; these are then used for learning prototypical representations of each class in semantic space, while the features describing tokens are based on the their distance, in semantic space, to each prototype vector. Providing the learning algorithm with these additional semantic features led to improved predictive performance. In Paper III, a manual inspection of the nearest neighbors of the prototype vectors revealed many desirable facts that are worth reiterating,

as they exemplify the idea underlying the method. First, almost all of the nearest neighbors seemed very reasonable and belonged to the corresponding named entity class. Second, many of the nearest neighbors were not present in the training data, demonstrating the potential of the semi-supervised method. Third, many of the nearest neighbors were misspellings or unusual spelling variants. Being able to capture these variants is crucial for successful named entity recognition in clinical text due to the prevalence of misspellings in this genre. Finally, the local neighborhoods of the prototype vectors inhabiting the various semantic spaces are rather different, producing largely independent rankings of nearest neighbors. This effectively indicates that there is, indeed, diversity across the semantic spaces in the ensemble.

Adverse drug event detection, addressed in Paper IV, V and VI, is approached in two rather different ways, with different purposes. In Paper IV and V, the problem is essentially cast as a binary classification tasks, detecting the presence or absence of a specific adverse drug event in a healthcare episode. The motivating use case is that some form of clinical decision support could be provided to improve the coding of adverse drug events in electronic health records. In the experimental investigations, heterogeneous clinical sequence data is model in ensembles of semantic spaces, resulting in improved predictive performance. In Paper VI, the task concerns relation extraction, detecting and classifying potential relations between named entity mentions in clinical notes that are assumed to be given a priori. Compared to the previous task, this one is more of basic component that could be used in a whole host of applications. This is a challenging task due to the large number of potential relations and the prevalence of long, inter-sentential relations. Here, the focus was on modeling the context of named entities – in the form of surrounding words – in ensembles of semantic space. For this particular task, using ensembles, as opposed to a single semantic space, did not always lead to improved predictive performance. Further analysis is required to understand why ensembles were less effective in this particular task.

CHAPTER 10

CONCLUSION

In this chapter, the work presented in the thesis is first briefly recapitulated. This is followed by a description and discussion of the most important contributions. Finally, a number of future directions are suggested.

10.1 RECAPITULATION

This thesis investigated the possibility of creating ensembles of semantic spaces to capture multiple aspects of the underlying data. The notions underpinning ensemble models were hence applied in a distributional semantic framework. In creating semantic space ensembles, a number of creation and combination strategies were explored: that is, ways of creating a set of potentially diverse semantic spaces and ways of combining them. The semantic space ensembles were empirically evaluated in tasks that have promising applications in healthcare: synonym extraction, named entity recognition and adverse drug event detection. In a number of experiments, it was demonstrated that semantic space ensembles often can – albeit not invariably so – outperform the use of a single semantic space.

Effective semantic space ensembles were created in various ways: (1) by manipulating the underlying data, more specifically by utilizing different types of corpora and different types of data; and (2) by manipulating various model hyperparameters, more specifically the size of the context window and the strategy for handling word order within that window. All of the investigated creation strategies yielded effective semantic space ensembles. Moreover, combinations of creation strategies yielded additional improvements in predictive performance.

A number of combination strategies were also compared. When the semantic space ensembles were used directly for k nearest neighbor retrieval, the most effective combination strategy, out of the ones considered, was to sum the cosine scores of the ranked lists produced by each semantic space and then returning a re-ranked list based on the summed scores. When the semantic space ensembles were used in a semi-supervised setting to generate features to be exploited by some learning algorithm, the most effective combination strategy was simply to fuse, through concatenation, the features generated by each semantic space.

A number of post-analyses were conducted in order to gain insights into the diversity of the semantic spaces in the created ensembles, as well as the contribution of each creation strategy to that end.

10.2 MAIN CONTRIBUTIONS

The single-most important contribution of this thesis is to have demonstrated the feasibility of combining distributional semantic spaces in order to create effective ensembles thereof. The criterion for effective ensembles is that they outperform, in the targeted task, any of the constituent models, in this case a single semantic space. In the empirical investigations, it was not only shown that this criterion is often met, but also that larger semantic space ensembles often outperform smaller ones, and that removing one type of semantic space – for instance, semantic spaces built with a certain window size or a certain word order strategy – tends to degrade predictive performance.

Furthermore, possible ways of creating and combining multiple semantic spaces that, to some degree, are diverse were proposed and evaluated. In this respect, emphasis was put on creation strategies, and two general approaches were investigated to that end: (1) it was shown that multiple corpora from different sub-domains can be exploited to create effective semantic space ensembles: these outperformed the use of a single semantic space built over the conjoined corpus, as well as multiple semantic spaces built over one of the corpora; (2) it was shown that certain hyperparameters of distributional semantic models, such as the size of the context window and the strategy employed for handling word order within that window, can be exploited to create effective semantic space ensembles: this allows different views of the same data to be created and effectively exploited to improve the predictive performance, compared to the use of a single semantic space, on a number of tasks.

Another important contribution is the extension of distributional semantics to non-linguistic sequence data. It was posited that the distributional hypothesis can, in fact, be extrapolated to other types of data, as long as they can meaningfully be represented sequentially. The extension of the distributional semantics framework to structured data allowed heterogeneous data to be modeled in semantic space, as well as in ensembles thereof, which has several advantages and readily allows these data types to be combined.

10.3 FUTURE DIRECTIONS

There are several future directions that could be taken based on the work presented in this thesis. The most obvious one is to explore alternative creation strategies, as well as to explore the proposed creation strategies in more depth. Beginning with the strategies involving manipulation of the underlying data, the notion of exploiting multiple types of corpora is in need of further study. One direction could be to exploit additional types of out-domain corpora, including large-scale corpora from the general domain. Another direction would instead be to investigate the division of an in-domain corpus into multiple sub-corpora with increasing specialization or into niche genres. In the medical domain, for instance, it would be possible to divide a corpus of clinical notes into many different corpora based on clinical practice. Rather than trying to capture the meaning of a word in general in a single representation, where the most frequent use of the word would dominate its representation, such an approach may allow the many nuances of a word to be better captured, in this case based on different uses in different clinical practices. This intuition, in fact, provides much of the motivation for multiple-corpora ensembles of semantic spaces, which goes beyond the fairly widespread endeavor of creating sense-specific word representations – in the sense that a specific number of senses can be defined or a priori or generally agreed upon – and views meaning as something inherently vague and multi-dimensional. Another direction with respect to manipulating the underlying data involves randomization techniques that have been successfully employed in other ensemble models such as random forest. One could, for instance, create bootstrap replicates of language use observations – defined, perhaps, as sentences – and/or randomly sample subsets of context features to create diverse semantic spaces. Other creation strategies could involve manipulating additional model hyperparameters – such as the actual context definition, different dimensionality reduction techniques, co-occurrence weighting strategies, distance measures, etc. – or using different distributional semantic models, e.g., a context-counting model such as random indexing and a context-predicting model such as skip-gram. With respect to creation strategies, it would also be interesting to explore whether sub-ensembles can be identified that are more effective than the original set of semantic spaces, as well as whether diversity can explicitly be optimized for in search of an effective semantic space ensemble.

Another potential direction would be to focus on combination strategies, as this aspect of ensembles was only investigated to a limited extent in this thesis. Many alternative strategies for combining ranked lists are conceivable in the case of k nearest neighbor retrieval; however, it may also be possible to combine vectors from multiple semantic spaces in some way prior to retrieval. As mentioned earlier, alternative strategies for fusing features generated by multiple semantic spaces and from different data sources could be explored. With respect to modeling heterogeneous data in semantic space, it would be interesting to try to model multiple data types in a shared space – or ensembles thereof – as this has been done successfully to create bilingual semantic spaces [Zou *et al.*, 2013] and to model images and words in a multi-modal semantic space [Socher *et al.*, 2013].

Finally, yet interesting direction would be to try to gain additional insights into the nature and success of semantic space ensembles, in particular what aspects of semantics are captured by various semantic spaces and (sub-)ensembles thereof. That would perhaps make it possible to create semantic space ensembles in a more principled and linguistically motivated manner.

REFERENCES

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgrén-Laine, Gunnar H Nilsson, Øystein Nytrø, *et al.* 2011. Characteristics of finnish and swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2(S-3):S1.

Ethern Alpaydin. 2010. *Introduction to Machine Learning*. MIT Press, Cambridge, MA. ISBN 978-0-262-01243-0.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

David Austen-Smith and Jeffrey S Banks. 1996. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review*, 90(01):34–45.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Association for Computational Linguistics*, volume 1, pages 238–247.

H.J.M. Beijer and C.J. De Blaey. 2002. Hospitalisations caused by adverse drug reactions (adr): a meta-analysis of observational studies. *Pharmacy World and Science*, 24(2):46–54.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55 (4):77–84.

Henrik Böstrom. 2007. Feature vs. classifier fusion for predictive data mining – a case study in pesticide classification. In *International Conference on Information Fusion*, pages 121–126. IEEE.

Henrik Boström, Sten F. Andler, Marcus Brohede, Ronnie Johansson, Er Karlsson, Joeri Van Laere, Lars Niklasson, Maria Nilsson, Anne Persson, and Tom Ziemke. 2007. On the definition of information fusion as a field of research. Technical report, Skövde: Institutionen för kommunikation och information.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Emmanuel Chazard, Gregoire Ficheur, Stephanie Bernonville, Michel Luyckx, and Regis Beuscart. 2011. Data mining to generate adverse drug events detection rules. *IEEE Transactions on Information Technology in Biomedicine*, 15(6):823–830.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Stephen Clark. 2012. Vector space models of lexical meaning. *Handbook of Contemporary Semantics, Wiley-Blackwell, à paraître*.

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

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

Preciosa M Coloma, Gianluca Trifirò, Vaishali Patadia, and Miriam Sturkenboom. 2013. Postmarketing safety surveillance. *Drug Safety*, 36(3):183–197.

Marquis de Condorcet. 1785. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In *Proceedings of the Swedish Language Technology Conference (SLTC)*.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus – Characteristics and Some Initial Findings. In *Proceedings of the 14th International Symposium on Health Information Management Research (ISHIMR)*, pages 1–7.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Scott C Deerwester, Susan T Dumais, Thomas K Landauer, George W Furnas, and Richard A Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Thomas G Dietterich. 2000a. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15. Springer-Verlag.

Thomas G Dietterich. 2000b. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157.

Robert Eriksson, Peter B Jensen, Sune Frankild, Lars J Jensen, and Søren Brunak. 2013a. Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. *Journal of the American Medical Informatics Association*.

Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. 2013b. Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. *Journal of the American Medical Informatics Association*, 20(5):947–953.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. In F. Palmer, editor, *Selected Papers of J. R. Firth 1952-59*, pages 168–205. Longman, London, UK.

Carol Friedman. 1997. Towards a comprehensive medical language processing system: methods and issues. In *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*, page 595. American Medical Informatics Association.

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

Curt D Furberg and Bertram Pitt. 2001. Withdrawal of cerivastatin from the world market. *Current Controlled Trials in Cardiovascular Medicine*, 2(5):205–207.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Stephen A Goldman. 1998. Limitations and strengths of spontaneous reports data. *Clinical Therapeutics*, 20:C40–C44.

Katja M Hakkarainen, Khadidja Hedna, Max Petzold, and Staffan Hägg. 2012. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions–a meta-analysis. *PloS One*, 7(3): e33236.

Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):993–1001.

Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. 2010. Mining electronic health records for adverse drug effects using regression based methods. In *the 1st ACM International Health Informatics Symposium*, pages 100–107. ACM.

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

Lorna Hazell and Saad AW Shakir. 2006. Under-reporting of adverse drug reactions. *Drug Safety*, 29(5):385–396.

Aron Henriksson. 2013. Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records.

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy W Chapman. 2013a. Identifying synonymy between snomed clinical terms of varying length using distributional analysis of electronic health records. In *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*, volume 2013, pages 600–608. American Medical Informatics Association.

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

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravičius, and Martin Hassel. 2012. 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)*.

Aron Henriksson, Maria Skeppstedt, Maria Kvist, Martin Duneld, and Mike Conway. 2013b. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In *Proceedings of BioNLP*. Association for Computational Linguistics.

Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

RL Howard, AJ Avery, S Slavenburg, S Royal, G Pipe, P Lucassen, and M Pirmohamed. 2007. Which drugs cause preventable admissions to hospital? a systematic review. *British Journal of Clinical Pharmacology*, 63(2):136–147.

Abhyuday N Jagannatha, Jinying Chen, and Hong Yu. 2015. Mining and ranking biomedical synonym candidates from wikipedia. In *Sixth International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 142–151.

Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

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

Michael N Jones and Douglas JK Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Jussi Karlgren, Anders Holst, and Magnus Sahlgren. 2008. Filaments of meaning in word space. In *Advances in Information Retrieval*, pages 531–538. Springer.

Isak Karlsson, Jing Zhao, Lars Asker, and Henrik Boström. 2013. Predicting adverse drug events by analyzing electronic patient records. In *Artificial Intelligence in Medicine Lecture Notes in Computer Science*, pages 125–129. Springer.

Samuel Kaski. 1998. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 1, pages 413–418. IEEE.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, pages 81–93.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pages 21–30.

Linda T Kohn, Janet M Corrigan, Molla S Donaldson, *et al.* 2000. *To Err Is Human:: Building a Safer Health System*, volume 627. National Academies Press.

Anders Krogh and Jesper Vedelsby. 1995. Neural network ensembles, cross validation, and cctive learning. In *Advances in Neural Information Processing Systems*, pages 231–238. MIT Press.

Ludmila I Kuncheva and Christopher J Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 160–170.

Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. 2013. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555.

Gondy Leroy and Hsinchun Chen. 2001. Meeting medical terminology needs-the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):261–270.

Gondy Leroy, James E Endicott, Obay Mouradi, David Kauchak, and Melissa L Just. 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings*, volume 2012, pages 522–531. American Medical Informatics Association.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

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

Dekang Lin and Patrick Pantel. 2001. Dirt - discovery of inference rules from text. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 323–328. ACM.

Hongfang Liu, Yves A Lussier, and Carol Friedman. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *Journal of Biomedical Informatics*, 34(4):249–261.

Tuwe Löfström. 2015. On effectively creating ensembles of classifiers: Studies on creation strategies, diversity and predicting with confidence.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic

health record: a review of recent research. *IMIA Yearbook of Medical Informatics*, 35:128–44.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Michael D Myers. 1997. Qualitative research in information systems. *Management Information Systems Quarterly*, 21:241–242.

Jonathan R Nebeker, Paul Barach, and Matthew H Samore. 2004. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Annals of Internal Medicine*, 140(10):795–801.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th conference on Computational linguistics*, volume 1, pages 304–309. Association for Computational Linguistics.

Oleg Okun and Giorgio Valentini. 2008. *Supervised and Unsupervised ensemble methods and their applications*. Springer. ISBN 978-3-540-78981-9.

David Opitz and Richard Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, pages 169–198.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM.

Bambang Parmanto, Paul W Munro, and Howard R Doyle. 1996. Improving committee diagnosis with resampling techniques. In *Advances in neural information processing systems*, pages 882–888.

Yves Peirsman, Kris Heylen, and Dirk Geeraerts. 2008. Size matters: tight and loose context definitions in english word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 34–41. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, 12:1532–1543.

Jordan B Pollack. 1990. Backpropagation is sensitive to initial conditions. *Complex Systems*, 4:269–80.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. *Semantic Evaluation*, 199(99):54.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Philip Resnik and Jimmy Lin. 2013. Evaluation of NLP Systems. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, West Sussex.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33 (1-2):1–39.

John I Saeed. 1997. *Semantics*. Blackwell Publishers, Oxford. ISBN 0-631-20034-7.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 1300–1305.

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

Steven L Salzberg. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3): 317–328.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Robert E Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5 (2):197–227.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proc. of EMNLP*.

Barbara Sibbald. 2004. Rofecoxib (vioxx) voluntarily withdrawn from market. *Canadian Medical Association Journal*, 171(9):1027–1028.

Maria Skeppstedt. 2014. *Extracting Clinical Findings from Swedish Health Record Text*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University.

Maria Skeppstedt, Magnus Ahltorp, and Aron Henriksson. 2013. Vocabulary expansion by semantic extraction of medical terms. In *The 5th International Symposium on Languages in Biology and Medicine (LBM 2013), Tokyo, Japan, 12-13 December, 2013*, pages 63–68.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Sveriges riksdag. 2013a. Sveriges riksdag: Lag (2003:460) om etikprövning av forskning som avser människor (In Swedish). URL http://www.riksdagen.se/sv/Dokument-Lagar/Ovriga-dokument/Ovrigt-dokument/_sfs-2003-460/. Accessed: September 30, 2013.

Sveriges riksdag. 2013b. Sveriges riksdag: Patientdatalag (2008:355) (In Swedish). URL http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Patientdatalag-2008355_sfs-2008-355/. Accessed: September 30, 2013.

Sveriges riksdag. 2013c. Sveriges riksdag: Personuppgiftslag (1998:204) (In Swedish). URL http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Personuppgiftslag-1998204_sfs-1998-204/. Accessed: September 30, 2013.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC Medical Informatics and Decision Making*, 13(Suppl 1):S1.

Peter D Turney. 2007. Empirical evaluation of four tensor decomposition algorithms. Technical report, Institute for Information Technology, National Research Council of Canada. Technical Report ERB-1152.

Peter D Turney and Michael L Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1): 141–188.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L. DuVall. 2011a. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of American Medical Informatics Association*, 18(5):552–556.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011b. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884.

Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28 (4):20.

Karin Wester, Anna K Jönsson, Olav Spigset, Henrik Druid, and Staffan Hägg. 2008. Incidence of fatal adverse drug reactions: a population based study. *British Journal of Clinical Pharmacology*, 65(4):573–579.

Dominic Widdows. 2004. *Geometry and Meaning*. CSLI Publications.

Ludwig Wittgenstein. 1958. *Philosophical investigations*. Blackwell Oxford.

David H Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

David H Wolpert. 1995. The relationship between pac, the statistical physics framework, the bayesian framework, and the vc framework. In *The Mathematics of Generalization*.

Lei Xu, Adam Krzyżak, and Ching Y Suen. 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435.

Jian Yang, Jing-yu Yang, David Zhang, and Jian-feng Lu. 2003. Feature fusion: parallel strategy vs. serial strategy. *Pattern Recognition*, 36(6):1369–1381.

G Udny Yule. 1900. On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pages 257–319.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. Uth_ccb: A report for semeval 2014–task 7 analysis of clinical text. *Semantic Evaluation 2014*, pages 802–806.

Jing Zhao, Aron Henriksson, Lars Asker, and Henrik Boström. 2014a. Detecting adverse drug events with multiple representations of clinical measurements. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 536–543. IEEE.

Jing Zhao, Aron Henriksson, and Henrik Boström. 2014b. Detecting adverse drug events using concept hierarchies of clinical codes. In *IEEE International Conference on Healthcare Informatics*, pages 285–293. IEEE.

Jing Zhao, Aron Henriksson, and Henrik Boström. 2015a. Cascading adverse drug event detection in electronic health records. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*. IEEE.

Jing Zhao, Aron Henriksson, Maria Kvist, Lars Asker, and Henrik Boström. 2015b. Handling temporality of clinical events for drug safety surveillance. In *American Medical Informatics Association (AMIA) Annual Symposium Proceedings*. American Medical Informatics Association.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1):239–263.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398.