Half Title

Title Page

LOC Page

To my dog and my cat.

Contents

Foreword	ix
Preface	xi
Contributors	xiii
Symbols	xvii
I This is What a Part Would Look Like	1
1 Natural Language Processing Author Name	3
1.0.1 Introduction to Natural Language Processing	4

Foreword

I am delighted to introduce the first book on Multimedia Data Mining. When I came to know about this book project undertaken by two of the most active young researchers in the field, I was pleased that this book is coming in early stage of a field that will need it more than most fields do. In most emerging research fields, a book can play a significant role in bringing some maturity to the field. Research fields advance through research papers. In research papers, however, only a limited perspective could be provided about the field, its application potential, and the techniques required and already developed in the field. A book gives such a chance. I liked the idea that there will be a book that will try to unify the field by bringing in disparate topics already available in several papers that are not easy to find and understand. I was supportive of this book project even before I had seen any material on it. The project was a brilliant and a bold idea by two active researchers. Now that I have it on my screen, it appears to be even a better idea.

Multimedia started gaining recognition in 1990s as a field. Processing, storage, communication, and capture and display technologies had advanced enough that researchers and technologists started building approaches to combine information in multiple types of signals such as audio, images, video, and text. Multimedia computing and communication techniques recognize correlated information in multiple sources as well as insufficiency of information in any individual source. By properly selecting sources to provide complementary information, such systems aspire, much like human perception system, to create a holistic picture of a situation using only partial information from separate sources.

Data mining is a direct outgrowth of progress in data storage and processing speeds. When it became possible to store large volume of data and run different statistical computations to explore all possible and even unlikely correlations among data, the field of data mining was born. Data mining allowed people to hypothesize relationships among data entities and explore support for those. This field has been put to applications in many diverse domains and keeps getting more applications. In fact many new fields are direct outgrowth of data mining and it is likely to become a powerful computational tool.

Preface

Approximately 17 million people in the USA (6% of the population) and 140 million people worldwide (this number is expected to rise to almost 300 million by the year 2025) suffer from diabetes mellitus. Currently, there a few dozens of commercialised devices for detecting blood glucose levels [1]. However, most of them are invasive. The development of a noninvasive method would considerably improve the quality of life for diabetic patients, facilitate their compliance for glucose monitoring, and reduce complications and mortality associated with this disease. Noninvasive and continuous monitoring of glucose concentration in blood and tissues is one of the most challenging and exciting applications of optics in medicine. The major difficulty in development and clinical application of optical noninvasive blood glucose sensors is associated with very low signal produced by glucose molecules. This results in low sensitivity and specificity of glucose monitoring by optical methods and needs a lot of efforts to overcome this difficulty.

A wide range of optical technologies have been designed in attempts to develop robust noninvasive methods for glucose sensing. The methods include infrared absorption, near-infrared scattering, Raman, fluorescent, and thermal gradient spectroscopies, as well as polarimetric, polarization heterodyning, photonic crystal, optoacoustic, optothermal, and optical coherence tomography (OCT) techniques [1-31].

For example, the polarimetric quantification of glucose is based on the phenomenon of optical rotatory dispersion, whereby a chiral molecule in an aqueous solution rotates the plane of linearly polarized light passing through the solution. The angle of rotation depends linearly on the concentration of the chiral species, the pathlength through the sample, and the molecule specific rotation. However, polarization sensitive optical technique makes it difficult to measure *in vivo* glucose concentration in blood through the skin because of the strong light scattering which causes light depolarization. For this reason, the anterior chamber of the eye has been suggested as a sight well suited for polarimetric measurements, since scattering in the eye is generally very low compared to that in other tissues, and a high correlation exists between the glucose in the blood and in the aqueous humor. The high accuracy of anterior eye chamber measurements is also due to the low concentration of optically active aqueous proteins within the aqueous humor.

On the other hand, the concept of noninvasive blood glucose sensing using the scattering properties of blood and tissues as an alternative to spectral absorption and polarization methods for monitoring of physiological glucose xii Preface

concentrations in diabetic patients has been under intensive discussion for the last decade. Many of the considered effects, such as changing of the size, refractive index, packing, and aggregation of RBC under glucose variation, are important for glucose monitoring in diabetic patients. Indeed, at physiological concentrations of glucose, ranging from 40 to 400 mg/dl, the role of some of the effects may be modified, and some other effects, such as glucose penetration inside the RBC and the followed hemoglobin glycation, may be important [30-32].

Noninvasive determination of glucose was attempted using light scattering of skin tissue components measured by a spatially-resolved diffuse reflectance or NIR frequency-domain reflectance techniques. Both approaches are based on change in glucose concentration, which affects the refractive index mismatch between the interstitial fluid and tissue fibers, and hence reduces scattering coefficient. A glucose clamp experiment showed that reduced scattering coefficient measured in the visible range qualitatively tracked changes in blood glucose concentration for the volunteer with diabetes studied.

Contributors

Michael Aftosmis

NASA Ames Research Center Moffett Field, California

Pratul K. Agarwal

Oak Ridge National Laboratory Oak Ridge, Tennessee

Sadaf R. Alam

Oak Ridge National Laboratory Oak Ridge, Tennessee

Gabrielle Allen

Louisiana State University Baton Rouge, Louisiana

Martin Sandve Alnæs

Simula Research Laboratory and University of Oslo, Norway Norway

Steven F. Ashby

Lawrence Livermore National Laboratory Livermore, California

David A. Bader

Georgia Institute of Technology Atlanta, Georgia

Benjamin Bergen

Los Alamos National Laboratory Los Alamos, New Mexico

Jonathan W. Berry

Sandia National Laboratories Albuquerque, New Mexico

Martin Berzins

University of Utah Salt Lake City, Utah

Abhinav Bhatele

University of Illinois Urbana-Champaign, Illinois

Christian Bischof

RWTH Aachen University Germany

Rupak Biswas

NASA Ames Research Center Moffett Field, California

Eric Bohm

University of Illinois Urbana-Champaign, Illinois

James Bordner

University of California, San Diego San Diego, California

George Bosilca

University of Tennessee Knoxville, Tennessee

Greg L. Bryan

Columbia University New York, New York

Marian Bubak

AGH University of Science and Technology Kraków, Poland xiv Contributors

Andrew Canning

Lawrence Berkeley National Laboratory Berkeley, California

Jonathan Carter

Lawrence Berkeley National Laboratory Berkeley, California

Zizhong Chen

Jacksonville State University Jacksonville, Alabama

Joseph R. Crobak

Rutgers, The State University of New Jersey Piscataway, New Jersey

Roxana E. Diaconescu

Yahoo! Inc. Burbank, California

Peter Diener

Louisiana State University Baton Rouge, Louisiana

Jack J. Dongarra

University of Tennessee, Knoxville, Oak Ridge National Laboratory, and University of Manchester

John B. Drake

Oak Ridge National Laboratory Oak Ridge, Tennessee

Kelvin K. Droegemeier

University of Oklahoma Norman, Oklahoma

Stéphane Ethier

Princeton University Princeton, New Jersey

Christoph Freundl

Friedrich–Alexander–Universität Erlangen, Germany

Karl Fürlinger

University of Tennessee Knoxville, Tennessee

Al Geist

Oak Ridge National Laboratory Oak Ridge, Tennessee

Michael Gerndt

Technische Universität München Munich, Germany

Tom Goodale

Louisiana State University Baton Rouge, Louisiana

Tobias Gradl

Friedrich-Alexander-Universität Erlangen, Germany

William D. Gropp

Argonne National Laboratory Argonne, Illinois

Robert Harkness

University of California, San Diego San Diego, California

Albert Hartono

Ohio State University Columbus, Ohio

Thomas C. Henderson

University of Utah Salt Lake City, Utah

Bruce A. Hendrickson

Sandia National Laboratories Albuquerque, New Mexico

Alfons G. Hoekstra

University of Amsterdam Amsterdam, The Netherlands

Philip W. Jones

Los Alamos National Laboratory Los Alamos, New Mexico Contributors

Laxmikant Kalé

University of Illinois Urbana-Champaign, Illinois

Shoaib Kamil

Lawrence Berkeley Berkeley, California

Cetin Kiris

NASA Ames Research Center Moffett Field, California

Uwe Küster

University of Stuttgart Stuttgart, Germany

Julien Langou

University of Colorado Denver, Colorado

Hans Petter Langtangen

Simula Research Laboratory and University of Oslo, Norway

Michael Lijewski

Lawrence Berkeley National Laboratory Berkeley, California

Anders Logg

Simula Research Laboratory and University of Oslo, Norway

Justin Luitjens

University of Utah Salt Lake City, Utah

Kamesh Madduri

Georgia Institute of Technology Atlanta, Georgia

Kent-Andre Mardal

Simula Research Laboratory and University of Oslo, Norway

Satoshi Matsuoka

Tokyo Institute of Technology Tokyo, Japan

John M. May

Lawrence Livermore National Laboratory Livermore, California

Celso L. Mendes

University of Illinois Urbana-Champaign, Illinois

Dieter an Mey

RWTH Aachen University Germany

Tetsu Narumi

Keio University Japan

Michael L. Norman

University of California, San Diego San Diego, California

Boyana Norris

Argonne National Laboratory Argonne, Illinois

Yousuke Ohno

Institute of Physical and Chemical Research (RIKEN) Kanagawa, Japan

Leonid Oliker

Lawrence Berkeley National Laboratory Berkeley, California

Brian O'Shea

Los Alamos of The National Laboratory Los Alamos, New Mexico

Christian D. Ott

University of Arizona Tucson, Arizona

James C. Phillips

University of Illinois Urbana-Champaign, Illinois xvi Contributors

Simon Portegies Zwart

University of Amsterdam, Amsterdam, The Netherlands

Thomas Radke

Albert-Einstein-Institut Golm, Germany

Michael Resch

University of Stuttgart Stuttgart, Germany

Daniel Reynolds

University of California, San Diego San Diego, California

Ulrich Rüde

Friedrich-Alexander-Universität Erlangen, Germany

Samuel Sarholz

RWTH Aachen University Germany

Erik Schnetter

Louisiana State University Baton Rouge, Louisiana

Klaus Schulten

University of Illinois Urbana-Champaign, Illinois

Edward Seidel

Louisiana State University Baton Rouge, Louisiana

John Shalf

Lawrence Berkeley National Laboratory Berkeley, California

Bo-Wen Shen

NASA Goddard Space Flight Center Greenbelt, Maryland

Ola Skavhaug

Simula Research Laboratory and University of Oslo, Norway

Peter M.A. Sloot

University of Amsterdam Amsterdam, The Netherlands

Erich Strohmaier

Lawrence Berkeley National Laboratory Berkeley, California

Makoto Taiji

Institute of Physical and Chemical Research (RIKEN) Kanagawa, Japan

Christian Terboven

RWTH Aachen University, Germany

Mariana Vertenstein

National Center for Atmospheric Research Boulder, Colorado

Rick Wagner

University of California, San Diego San Diego, California

Daniel Weber

University of Oklahoma Norman, Oklahoma

James B. White, III

Oak Ridge National Laboratory Oak Ridge, Tennessee

Terry Wilmarth

University of Illinois Urbana-Champaign, Illinois

Symbols

Symbol Description

α	To solve the generator maintenance scheduling, in the		annealing and genetic algorithms have also been tested.
	9,	$\theta \sqrt{abc}$	This paper presents a survey
	techniques have been ap-		of the literature
	plied.	ζ	over the past fifteen years in
σ^2	These include integer pro-		the generator
	gramming, integer linear	∂	maintenance scheduling.
	programming, dynamic pro-		The objective is to
	gramming, branch and	sdf	present a clear picture of the
	bound etc.		available recent literature
\sum	Several heuristic search al-	ewq	of the problem, the con-
	gorithms have also been de-		straints and the other as-
	veloped. In recent years ex-		pects of
	pert systems,	bvcn	the generator maintenance
abc	fuzzy approaches, simulated		schedule.

Part I This is What a Part Would Look Like

Natural Language Processing

Author Name

CONTENTS

1.0.1 Introduction to Natural Language Processing

In this chapter we introduce modern NLP libraries, techniques and their applications. This chapter will focus on deep learning methods and less on computational linguistics that require nuanced knowledge of linguistics. We explore what it means to represent words and sequences of words with rich numeric representations that are better-suited toward modern computational tasks. We aim to capture some of these modern fine-tuned representations that are specially catered toward a semantic lexicon for medical language. We use these representations and aforementioned tools to showcase a modern reference implementation leveraging PyTorch, PyTorch Lightning and the Hugginface Transformers library. To wrap it all together, we walk through a complete example that highlights best practices that encourage reproducibility and allow for systematic iterative improvements.

This includes:

- Bootstrapping techniques to iterate on a dataset in the low-resource setting
- Storing of a reference dataset in a publicly-accessible location
- Downloading, caching, loading, splitting, and preprocessing of the data
- Setting up of a cloud-based GPU workstation (?) (-this might be overkill for now, but keep if we can)
- VSCode (?)
- Monitoring the training run:

Logging and experiment tracking

Learning curves

Metrics

- Hyperparameter tuning, some tricks of the trade
- Offline evaluation and sanity checking

We will keep the discussion focused on SUDEP prediction from electronic medical record (EMR) notes. Many of the concepts introduced here are very general and are straightforward translations to domains outside of SUDEP prediction, epilepsy, and even NLP.

1.0.1 Introduction to Natural Language Processing

Natural language processing (NLP) is a field of computer science that deals with the extraction, processing, and understanding of human language. It is known as the field of computer linguistics, and is a subfield of artificial intelligence. Common NLP tasks include sentence segmentation, tokenization, part-of-speech tagging, named-entity recognition, parsing, question answering, summarization and classification. How can we teach a computer to perform these tasks? The first challenge is that computers at their core only understand numbers, so we need to represent words with numbers. To do this we can define a vocabulary V of words and numbers, where each word is assigned a unique integer i. The word is then represented as a vector w of length |V| with all zeros and a one at index i. We could now simply represent a sentence as the sum of the word vectors $S = \sum_j w_j$. This is a simple representation and it comes with some drawbacks:

- We implicitly assume that each word is equally important.
- Each word is equally similar to every other word (e.g. by taking the euclidian distance).
- The representation is invariant to reordering of the words.

To address the first point we can instead write a weighted sum $S = \sum_j \lambda_j w_j$, where each word is weighted by its importance λ_j . One way of defining importance is to compare how often a word occurs in a document compared to the entire corpus.