Explainable
Natural Language Processing

Synthesis Lectures on Human Language Technologies

Synthesis Lectures on Human Language Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Explainable Natural Language Processing
Anders Søgaard
2021

Finite-State Text Processing
Kyle Gorman and Richard Sproat
2021

Semantic Relations Between Nominals, Second Edition
Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séagdha
2021

Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning
Mohammad Taher Pilehvar and Jose Camacho-Collados
2020

Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots
Michael McTear
2020

Natural Language Processing for Social Media, Third Edition
Anna Atefeh Farzindar and Diana Inkpen
2020

Statistical Significance Testing for Natural Language Processing
Rotem Dror, Lotem Peled, Segev Shlomov, and Roi Reichart
2020

Deep Learning Approaches to Text Production
Shashi Narayan and Claire Gardent
2020

Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics
Emily M. Bender and Alex Lascarides
2019

Cross-Lingual Word Embeddings
Anders Søgaard, Ivan Vulić, Sebastian Ruder, Manaal Faruqui
2019

Bayesian Analysis in Natural Language Processing, Second Edition
Shay Cohen
2019

Argumentation Mining
Manfred Stede and Jodi Schneider
2018

Quality Estimation for Machine Translation
Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold
2018

Natural Language Processing for Social Media, Second Edition
Atefeh Farzindar and Diana Inkpen
2017

Automatic Text Simplification
Horacio Saggion
2017

Neural Network Methods for Natural Language Processing
Yoav Goldberg
2017

Syntax-based Statistical Machine Translation
Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn
2016

Domain-Sensitive Temporal Tagging
Jannik Strötgen and Michael Gertz
2016

Linked Lexical Knowledge Bases: Foundations and Applications
Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek
2016

Bayesian Analysis in Natural Language Processing
Shay Cohen
2016

Metaphor: A Computational Perspective
Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov
2016

Grammatical Inference for Computational Linguistics
Jeffrey Heinz, Colin de la Higuera, and Menno van Zaanen
2015

Automatic Detection of Verbal Deception
Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari
2015

Natural Language Processing for Social Media
Atefeh Farzindar and Diana Inkpen
2015

Semantic Similarity from Natural Language and Ontology Analysis
Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
2015

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition
Hang Li
2014

Ontology-Based Interpretation of Natural Language
Philipp Cimiano, Christina Unger, and John McCrae
2014

Automated Grammatical Error Detection for Language Learners, Second Edition
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2014

Web Corpus Construction
Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications
Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Explainable Natural Language Processing

Anders Søgaard

ISBN: 978-3-031-01052-1 paperback
ISBN: 978-3-031-02180-0 ebook
ISBN: 978-3-031-00191-8 hardcover

DOI 10.1007/978-3-031-02180-0

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #51
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Explainable Natural Language Processing

Anders Søgaard
University of Copenhagen

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #51

ABSTRACT

This book presents a taxonomy framework and survey of methods relevant to explaining the decisions and analyzing the inner workings of Natural Language Processing (NLP) models. The book is intended to provide a snapshot of Explainable NLP, though the field continues to rapidly grow. The book is intended to be both readable by first-year M.Sc. students and interesting to an expert audience. The book opens by motivating a focus on providing a consistent taxonomy, pointing out inconsistencies and redundancies in previous taxonomies. It goes on to present (i) a taxonomy or framework for thinking about how approaches to explainable NLP relate to one another; (ii) brief surveys of each of the classes in the taxonomy, with a focus on methods that are relevant for NLP; and (iii) a discussion of the inherent limitations of some classes of methods, as well as how to best evaluate them. Finally, the book closes by providing a list of resources for further research on explainability.

KEYWORDS

natural language processing, interpretability, explainability, taxonomies, resources

Contents

Acknowledgments . . . . **xv**

1 Introduction **1**

1.1 Two Common Distinctions 3

1.1.1 Local and Global Explanations 3

1.1.2 Intrinsic and Post-Hoc Explanations 5

1.2 Shortcomings of Existing Taxonomies 6

1.2.1 Guidotti et al. (2018) 6

1.2.2 Adadi and Berrada (2019) 7

1.2.3 Carvalho et al. (2019) 7

1.2.4 Molnar (2019) 8

1.2.5 Zhang et al. (2020) 8

1.2.6 Danilevsky et al. (2020) 9

1.2.7 Das et al. (2020) 9

1.2.8 Atanasova et al. (2020) 10

1.2.9 Kotonya and Toni (2020) 10

1.3 The Method-Form Fallacy 11

1.4 Inconsistent Classifications 11

1.5 A Novel Taxonomy 11

2 A Framework for Explainable NLP **13**

2.1 NLP Architectures 13

2.1.1 Linear and Nonlinear Classification 13

2.1.2 Recurrent Models 14

2.1.3 Transformers 15

2.1.4 Overview of Applications and Architectures 16

2.2 Local and Global Explanations 17

2.3 Backward Methods 17

2.4 Forward Explaining by Intermediate Representations 18

2.5 Forward Explaining by Continuous Outputs 19

2.6 Forward Explaining by Discrete Outputs 19

3 Local-Backward Explanations 21
3.1 Vanilla Gradients 21
3.2 Guided Back-Propagation 21
3.3 Layer-Wise Relevance Propagation 22
3.4 Deep Taylor Decomposition 23
3.5 Integrated Gradients 23
3.6 DeepLift 24

4 Global-Backward Explanations 25
4.1 Post-Hoc Unstructured Pruning 25
4.2 Lottery Tickets 26
4.3 Dynamic Sparse Training 27
4.4 Binary Networks and Sparse Coding 28

5 Local-Forward Explanations of Intermediate Representations 31
5.1 Gates 31
5.2 Attention 31
5.3 Attention Roll-Out and Attention Flow 32
5.4 Layer-Wise Attention Tracing 33
5.5 Attention Decoding 34

6 Global-Forward Explanations of Intermediate Representations 35
6.1 Gate Pruning 35
6.2 Attention Head Pruning 36

7 Local-Forward Explanations of Continuous Output 39
7.1 Word Association Norms 39
7.2 Word Analogies 39
7.3 Time Step Dynamics 40

8 Global-Forward Explanations of Continuous Output 43
8.1 Correlation of Representations 43
8.2 Clustering 44
8.3 Probing Classifiers 44
8.4 Concept Activation 45
8.5 Influential Examples 46

9 Local-Forward Explanations of Discrete Output 49

9.1 Challenge Datasets 49

9.2 Local Uptraining 50

9.3 Influential Examples 51

10 Global-Forward Explanations of Discrete Output 53

10.1 Uptraining 53

10.2 Meta-Analysis 54

10.3 Downstream Evaluation 54

11 Evaluating Explanations 55

11.1 Flavors of Explanations 55

11.2 Heuristics 58

11.3 Human Annotations 60

11.4 Human Experiments 61

12 Perspectives 63

12.1 General Observations 63

12.2 Beyond Taxonomy 67

12.3 Moral Foundations of Explanations 68

13 Resources 71

13.1 Code 71

13.2 Datasets and Benchmarks 73

Bibliography 75

Author's Biography 107

Acknowledgments

The book's shortcomings are entirely my fault. I would nevertheless like to thank my editors and reviewers, as well as several of my colleagues for engaging in discussions that helped shape this book: Mostafa Abdou, Jasmijn Bastings, Katja Filippova, Ana Valeria Gonzalez, Daniel Herschcovich, Karthikeyan K, Anna Rogers, and Anna Rumshinsky. The framework presented here is probably applicable outside of NLP, but I willfully ignore all other applications of interpretability and explainable machine learning throughout the book.

Anders Søgaard
September 2021