



Cognitive Approach to Natural Language Processing

Edited by Bernadette Sharp

Florence Sèdes and Wiesław Lubaszewski

**ISTE
PRESS**



As natural language processing spans many different disciplines, it is sometimes difficult to understand the contributions and the challenges that each of them presents. This book explores the special relationship between natural language processing and cognitive science, and the contribution of computer science to these two fields. It is based on the recent research papers submitted at the international workshops of Natural Language and Cognitive Science (NLPCS) which was launched in 2004 in an effort to bring together natural language researchers, computer scientists, and cognitive and linguistic scientists to collaborate together and advance research in natural language processing.

The chapters cover areas related to language understanding, language generation, word association, word sense disambiguation, word predictability, text production and authorship attribution. This book will be relevant to students and researchers interested in the interdisciplinary nature of language processing.

Bernadette Sharp is Professor of Applied Artificial Intelligence (AI) at Staffordshire University, UK. Her research interests include AI, natural language processing, and text mining. She has been Chair and Editor of the International Workshop for Natural Language Processing and Cognitive Science since 2004.

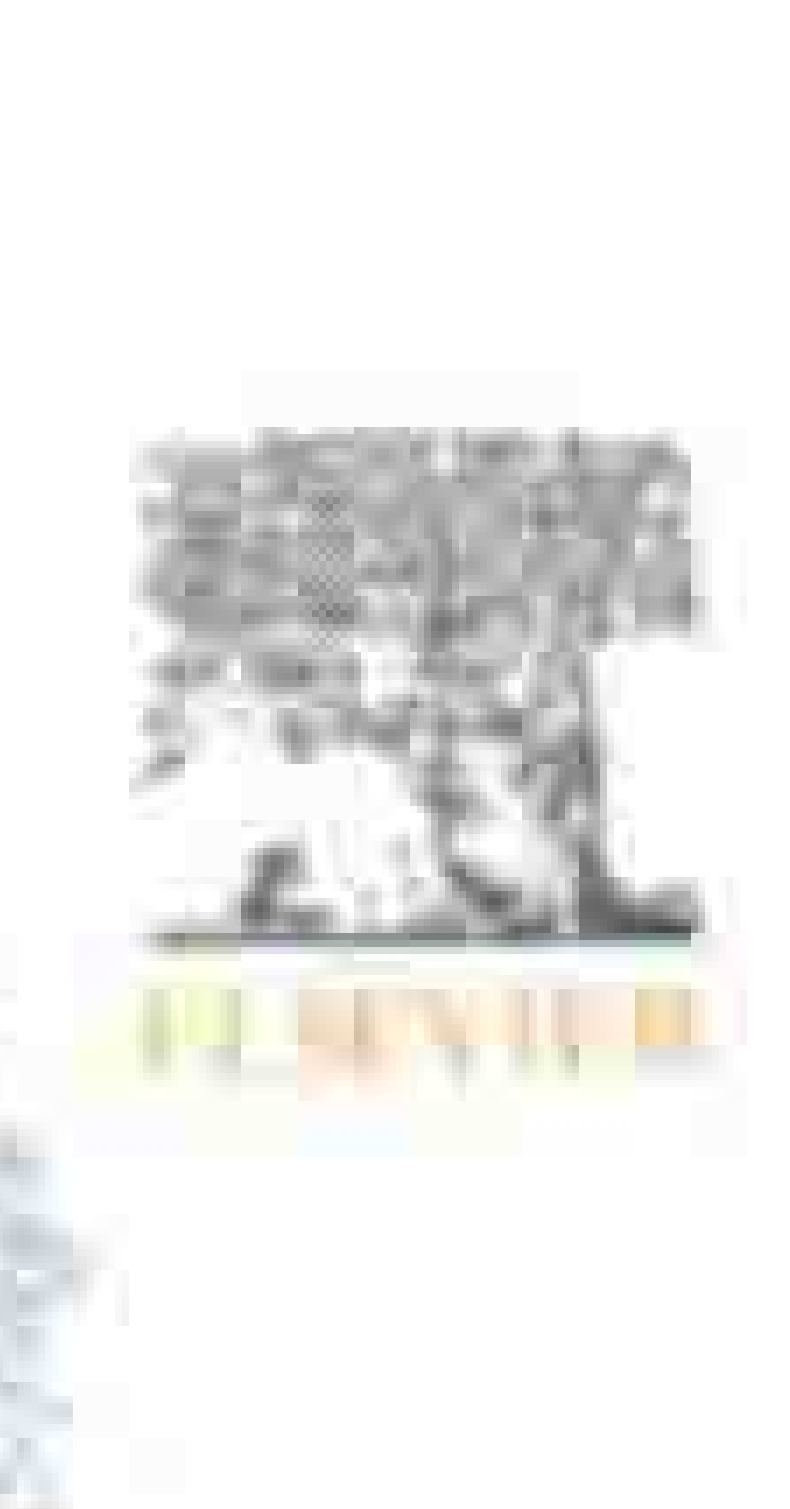
Florence Sèdes is Professor of Computer Science at Toulouse University, France. Her research areas cover information systems and data management with applications dedicated to multimedia, metadata and mobility in ambient intelligence, social media and CCTV. She supervises a “smart restaurant” platform for emotion and social interaction analysis, and contributes to the ISO 22311 standard.

Wiesław Lubaszewski is Professor at the Department of Computational Linguistics of the Jagiellonian University and Professor at the Computer Science Department of AGH, University of Technology, in Kraków, Poland. His research interests include natural language dictionaries, text understanding, knowledge representation, and information extraction.



Edited by B. Sharp
F. Sedes, W. Lubaszewski

Cognitive Approach to Natural Language Processing



Series Editor
Florence Sèdes

Cognitive Approach to Natural Language Processing

Edited by

Bernadette Sharp
Florence Sèdes
Wiesław Lubaszewski

ISTE
PRESS



First published 2017 in Great Britain and the United States by ISTE Press Ltd and Elsevier Ltd

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Press Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

Elsevier Ltd
The Boulevard, Langford Lane
Kidlington, Oxford, OX5 1GB
UK

www.elsevier.com

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

For information on all our publications visit our website at <http://store.elsevier.com/>

© ISTE Press Ltd 2017

The rights of Bernadette Sharp, Florence Sèdes and Wiesław Lubaszewski to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

A catalog record for this book is available from the Library of Congress

ISBN 978-1-78548-253-3

Cognitive Approach to Natural Language Processing

Preface

This book is a special issue dedicated to exploring the relationship between natural language processing and cognitive science, and the contribution of computer science to these two fields. Poibeau and Vasishth [POI 16] noted that research interest in cognitive issues may have been given less attention because researchers from the cognitive science field are overwhelmed by the technical complexity of natural language processing; similarly, natural language processing researchers have not recognized the contribution of cognitive science to their work. We believe that the international workshops of Natural Language and Cognitive Science (NLPCS), launched in 2004, have provided a strong platform to support the consistent determination and diversity of new research projects which acknowledge the importance of interdisciplinary approaches and bring together computer scientists, cognitive and linguistic researchers to advance research in natural language processing.

This book consists of 10 chapters contributed by the researchers at the recent NLPCS workshops. In Chapter 1, Philippe Blache explains that the process of understanding language is theoretically very complex; it must be carried out in real time. This process requires many different sources of information. He argues that the global interpretation of a linguistic input is based on the grouping of elementary units called chunks which constitute the backbone of the “interpret whenever possible” principle which is responsible for delaying the understanding process until enough information becomes available. The following two chapters address the problem of human association. In Chapter 2, Korzycki, Gatkowska and Lubaszewski discuss an experiment based on 900 students who participated in a free word

association test. They have compared the human association list with the association list retrieved from text using three algorithms: the Church and Hanks algorithm, the Latent Semantic Analysis and Latent Dirichlet Allocation. In Chapter 3, Lubaszewski, Gatkowska and Godny describe a procedure developed to investigate word associations in an experimentally built human association network. They argue that each association is based on the semantic relation between two meanings, which has its own direction and is independent from the direction of other associations. This procedure uses graph structures to produce a semantically consistent subgraph. In Chapter 4, Rapp investigates whether human language generation is governed by associations, and whether the next content word of an utterance can be considered as an association with the representations of the content words, already activated in the speaker's memory. He introduces the concept of the Reverse Association Task and discusses whether the stimulus can be predicted from the responses. He has collected human data based on the reverse association task, and compared them to the machine-generated results. In Chapter 5, Vincent-Lamarre and his colleagues have investigated how many words, and which ones, are required to define all the rest of the words in a dictionary. To this end, they have applied graph-theoretic analysis to the Wordsmyth suite of dictionaries. The results of their study have implications for the understanding of symbol grounding and the learning and mental representation of word meaning. They conclude that language users must have the vocabulary to understand the words in definitions to be able to learn and understand the meaning of words from verbal definitions. Chapter 6 focuses on word sense disambiguation. Tripodi and Pelillo have explored the evolutionary game theory approach to study word sense disambiguation. Each word to be disambiguated is represented as a player and each sense as a strategy. The algorithm has been tested on four datasets with different numbers of labeled words. It exploits relational and contextual information to infer the meaning of a target word. The experimental results demonstrate that this approach has outperformed conventional methods and requires a small amount of labeled points to outperform supervised systems. In Chapter 7, Zock and Tesfaye have focused on the challenging task of text production expressed in terms of four tasks: ideation, text structuring, expression and revision. They have focused on text structuring which involves the grouping (chunking), ordering and linking of messages. Their aim is to study which parts of text production can be automated, and whether the computer can build one or several topic trees based on a set of inputs provided by the user. Authorship attribution is the focus of study in

Chapter 8. Boukhaled and Ganascia have analyzed the effectiveness of using sequential rules of function words and Part-of-Speech (POS) tags as a style marker that does not rely on the bag-of-words assumption or on their raw frequencies. Their study has shown that the frequencies of function words and POS n-grams outperform the sequential rules. Fundamental frequency detection (F0), which plays an important role in human speech perception, is addressed in Chapter 9. Glavitsch has investigated whether F0 estimation, using the principles of human cognition, can perform equally well or better than state-of-the-art F0 detection algorithms. The proposed algorithm, which operates in the time domain, has achieved very low error rates and outperformed the state-of-the-art correlation-based method RAPT in this respect, using limited resources in terms of memory and computing power. In neurocognitive psychology, manually collected cloze completion probabilities (CCPs) are used to quantify the predictability of a word from sentence context in models of eye movement control. As these CCPs are based on samples of up to 100 participants, it is difficult to generalize a model across all novel stimuli. In Chapter 10, Hofmann, Biemann and Remus have proposed applying language models which can be benchmarked by item-level performance on datasets openly available in online databases. Previous neurocognitive approaches to word predictability from sentence context in electroencephalographic (EEG) and eye movement (EM) data relied on cloze completion probability (CCP) data. Their study has demonstrated that the syntactic and short-range semantic processes of n-gram language models and recurrent neural networks (RNN) can perform more or less equally well when directly accounting CCP, EEG and EM data. This may help generalize neurocognitive models to all possible novel word combinations.

Bibliography

- [POI 16] POIBEAU T., VASISHTH S., “Introduction: Cognitive Issues in Natural Language Processing”, *Traitement Automatique des Langues et Sciences Cognitives*, vol. 55, no. 3, pp. 7–19, 2016.

Bernadette SHARP
 Florence SÈDES
 Wiesław LUBASZEWSKI
 March 2017

Contents

Preface	xi
--------------------------	----

Chapter 1. Delayed Interpretation, Shallow Processing and Constructions: the Basis of the “<i>Interpret Whenever Possible</i>” Principle	1
---	---

Philippe BLACHE

1.1. Introduction.	1
1.2. Delayed processing	3
1.3. Working memory	5
1.4. How to recognize chunks: the segmentation operations	8
1.5. The delaying architecture.	10
1.5.1. Segment-and-store	11
1.5.2. Aggregating by cohesion	12
1.6. Conclusion	16
1.7. Bibliography	17

Chapter 2. Can the Human Association Norm Evaluate Machine-Made Association Lists?	21
---	----

Michał KORZYCKI, Izabela GATKOWSKA
and Wiesław LUBASZEWSKI

2.1. Introduction.	21
2.2. Human semantic associations	23
2.2.1. Word association test.	23
2.2.2. The author’s experiment.	24
2.2.3. Human association topology	25
2.2.4. Human associations are comparable.	26

2.3. Algorithm efficiency comparison	29
2.3.1. The corpora	29
2.3.2. LSA-sourced association lists.	29
2.3.3. LDA-sourced lists	31
2.3.4. Association ratio-based lists	31
2.3.5. List comparison	32
2.4. Conclusion	37
2.5. Bibliography	38

**Chapter 3. How a Word of a Text Selects the
Related Words in a Human Association Network 41**

Wiesław LUBASZEWSKI, Izabela GATKOWSKA
and Maciej GODNY

3.1. Introduction.	41
3.2. The network	44
3.3. The network extraction driven by a text-based stimulus	46
3.3.1. Sub-graph extraction algorithm.	46
3.3.2. The control procedure	48
3.3.3. The shortest path extraction.	48
3.3.4. A corpus-based sub-graph.	50
3.4. Tests of the network extracting procedure.	50
3.4.1. The corpus to perform tests	50
3.4.2. Evaluation of the extracted sub-graph.	51
3.4.3. Directed and undirected sub-graph extraction: the comparison	52
3.4.4. Results per stimulus	53
3.5. A brief discussion of the results and the related work	58
3.6. Bibliography	60

Chapter 4. The Reverse Association Task 63

Reinhard RAPP

4.1. Introduction.	63
4.2. Computing forward associations	67
4.2.1. Procedure.	67
4.2.2. Results and evaluation	69
4.3. Computing reverse associations.	71
4.3.1. Problem.	71
4.3.2. Procedure.	71
4.3.3. Results and evaluation	76

4.4. Human performance	78
4.4.1. Dataset	78
4.4.2. Test procedure.	80
4.4.3. Evaluation	81
4.5. Performance by machine	82
4.6. Discussion, conclusions and outlook	84
4.6.1. Reverse associations by a human	84
4.6.2. Reverse associations by a machine	85
4.7. Acknowledgments.	87
4.8. Bibliography	88

Chapter 5. Hidden Structure and Function in the Lexicon 91

Philippe VINCENT-LAMARRE, Mélanie LORD,
 Alexandre BLONDIN-MASSÉ, Odile MARCOTTE,
 Marcos LOPES and Stevan HARNAD

5.1. Introduction.	91
5.2. Methods	92
5.2.1. Dictionary graphs.	92
5.2.2. Psycholinguistic variables.	96
5.2.3. Data analysis.	96
5.3. Psycholinguistic properties of Kernel, Satellites, Core, MinSets and the rest of each dictionary	97
5.4. Discussion	101
5.4.1. Limitations.	104
5.5. Future work.	104
5.6. Bibliography	106

Chapter 6. Transductive Learning Games for Word Sense Disambiguation 109

Rocco TRIPODI and Marcello PELILLO

6.1. Introduction.	109
6.2. Graph-based word sense disambiguation	111
6.3. Our approach to semi-supervised learning	113
6.3.1. Graph-based semi-supervised learning	113
6.3.2. Game theory and game dynamics	114
6.4. Word sense disambiguation games	116
6.4.1. Graph construction	116
6.4.2. Strategy space	117
6.4.3. The payoff matrix.	118
6.4.4. System dynamics	119

6.5. Evaluation.	120
6.5.1. Experimental setting	120
6.5.2. Evaluation results.	121
6.5.3. Comparison with state-of-the-art algorithms.	124
6.6. Conclusion	124
6.7. Bibliography	125

Chapter 7. Use Your Mind and Learn to Write: The Problem of Producing Coherent Text 129

Michael ZOCK and Debela Tesfaye GEMECHU

7.1. The problem	129
7.2. Suboptimal texts and some of the reasons.	131
7.2.1. Lack of coherence or cohesion	132
7.2.2. Faulty reference.	133
7.2.3. Unmotivated topic shift	134
7.3. How to deal with the complexity of the task?.	135
7.4. Related work	136
7.5. Assumptions concerning the building of a tool assisting the writing process.	138
7.6. Methodology	141
7.6.1. Identification of the syntactic structure	143
7.6.2. Identification of the semantic seed words.	144
7.6.3. Word alignment.	145
7.6.4. Determination of the similarity values of the aligned words.	146
7.6.5. Determination of the similarity between sentences	150
7.6.6. Sentence clustering based on their similarity values	151
7.7. Experiment and evaluation.	151
7.8. Outlook and conclusion.	154
7.9. Bibliography	155

Chapter 8. Stylistic Features Based on Sequential Rule Mining for Authorship Attribution 159

Mohamed Amine BOUKHALED and Jean-Gabriel GANASCIA

8.1. Introduction and motivation	159
8.2. The authorship attribution process	162
8.3. Stylistic features for authorship attribution	163
8.4. Sequential data mining for stylistic analysis	165

8.5. Experimental setup	166
8.5.1. Dataset	166
8.5.2. Classification scheme	167
8.6. Results and discussion	169
8.7. Conclusion	173
8.8. Bibliography	173

Chapter 9. A Parallel, Cognition-oriented Fundamental Frequency Estimation Algorithm 177

Ulrike GLAVITSCH

9.1. Introduction.	177
9.2. Segmentation of the speech signal	180
9.2.1. Speech and pause segments	180
9.2.2. Voiced and unvoiced regions	182
9.2.3. Stable and unstable intervals	183
9.3. F0 estimation for stable intervals	184
9.4. F0 propagation	186
9.4.1. Control flow	187
9.4.2. Peak propagation	189
9.5. Unstable voiced regions	191
9.6. Parallelization	191
9.7. Experiments and results	192
9.8. Conclusions.	194
9.9. Acknowledgments.	195
9.10. Bibliography	195

Chapter 10. Benchmarking n-grams, Topic Models and Recurrent Neural Networks by Cloze Completions, EEGs and Eye Movements 197

Markus J. HOFMANN, Chris BIEMANN and Steffen REMUS

10.1. Introduction	198
10.2. Related work	199
10.3. Methodology	200
10.3.1. Human performance measures	200
10.3.2. Three flavors of language models	201
10.4. Experiment setup.	203
10.5. Results	204
10.5.1. Predictability results	204
10.5.2. N400 amplitude results.	206
10.5.3. Single-fixation duration (SFD) results.	208

10.6. Discussion and conclusion	210
10.7. Acknowledgments	212
10.8. Bibliography	212
List of Authors	217
Index	219