Cognitive Approach to Natural Language Processing

Edited by Bernadette Sharp
Florence Sèdes and Wiesław Lubaszewski

iSTE PRESS
ELSEVIER

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.

ELSEVIER

iSTE PRESS

www.iste.co.uk

9 781785 482533

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
ELSEVIER

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

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.

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