

Studies in Computational Intelligence

Volume 918

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are submitted to indexing to Web of Science, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink.

More information about this series at <http://www.springer.com/series/7092>

Chalermpol Tapsai · Herwig Unger ·
Phayung Meesad

Thai Natural Language Processing

Word Segmentation, Semantic Analysis,
and Application

Chalermpol Tapsai
College of Innovation and Management
Suan Sunandha Rajabhat University
Bangkok, Thailand

Herwig Unger
Fakultät für Mathematik und Informatik
FernUniversität in Hagen
Hagen, Nordrhein-Westfalen, Germany

Phayung Meesad
Faculty of Information Technology
and Digital Information, Department
of Information Technology Management
King Mongkut's University of Technology
North Bangkok
Bangkok, Thailand

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-030-56234-2 ISBN 978-3-030-56235-9 (eBook)
<https://doi.org/10.1007/978-3-030-56235-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The purpose of this book is to present a complete practical process of Thai Natural Language Processing (NLP) which is suitable for both new NLP developers and those who already have experience. By composing the content step by step, the new developers can follow each chapter to complete all NLP processes with both theory and practical workshops. New techniques, such as Ranking Trie and Completed Soundex, are presented to give the idea for solving crucial problems in NLP. For clear reading content, Chap. 1 presents the fundamentals of NLP steps and Thai language principles for the essential background that help readers understand the content of Chaps. 2–4 more clearly. Chapter 5 is the NLP programming workshop in both English and Thai. The last chapter presents an example of complete NLP research, which covers all processing steps that help readers get an idea for further application at a higher level.

The authors want to thank our colleagues for their supports as well as assistance from staff that facilitates our operations in all steps.

Bangkok, Thailand
Hagen, Germany
Bangkok, Thailand

Chalermopol Tapsai
Herwig Unger
Phayung Meesad

Contents

1	Introduction	1
1.1	Natural Language Processing	1
1.2	Fundamental Knowledge of Thai Language Principles	4
1.2.1	Thai Alphabets	5
1.2.2	Thai Word	11
1.2.3	Transforming of Verbs into Nouns	14
1.2.4	Transforming of Adjectives into Nouns	15
1.2.5	Transforming of Adjectives into Adverbs	16
1.2.6	Numeral and Quantity Representation	16
1.2.7	Quantifying Noun	17
1.3	Thai Sentences	19
	References	23
2	Thai Word Segmentation	25
2.1	Syllable Segmentation	25
2.2	Word Segmentation	27
2.3	Trie Is Not Tree	27
2.4	Word Segmentation Based on Surrounding Contexts	33
2.5	Comparison of Thai Segmentation Algorithms	34
2.6	Problems in Thai Word Segmentation	35
	References	36
3	TLS-ART-MC, A New Algorithm for Thai Word Segmentation	37
3.1	The TLS-ART-MC Algorithm	37
3.2	Datasets	39
3.3	Model Development and Evaluation	39
3.4	Ranking Trie	39
3.4.1	Ranking Trie Creation Algorithm	41
3.5	Word Usage Frequency Analysis	43

3.5.1	Text Corpus	43
3.5.2	The Results of Word Usage Frequency Analysis	44
3.5.3	Character Statistics	44
3.5.4	Consonants and Vowels	48
3.5.5	Word Types	49
3.6	Word Segmentation with Automatic Ranking Trie	50
3.7	Solving Problems of Misspelling and Various Spelling Patterns	52
3.7.1	Soundex	53
3.7.2	Traditional Soundex Code	53
3.7.3	Completed Soundex	63
3.7.4	Completed Soundex Encoding	66
3.7.5	Completed Soundex Encoding Process	70
3.7.6	Completed Soundex Similarity Values	74
3.7.7	Evaluation of Completed Soundex	76
3.7.8	The Experiment for Performance Evaluation	77
3.8	Conclusion	82
	References	83
4	Semantic Analysis	85
4.1	Pattern Parsing	85
4.2	Ontology	87
4.3	Semantic Pattern	94
4.4	Summarization	96
	References	96
5	Thai Natural Language Processing Programming	99
5.1	NLP Programming Tools	99
5.2	Basic Programming with Python	104
5.3	Control Statements	113
5.3.1	If Statements	113
5.3.2	Nested-If Statements	118
5.4	Loop Statements	119
5.4.1	While Statements	119
5.4.2	For Statements	120
5.5	Workshop: Development of NLP Program (English)	121
5.6	Workshop: Development of NLP Program (Thai)	127
5.7	Summarization	130
	References	130
6	The Application of Thai Natural Language Processing	131
6.1	Conceptual Framework	131
6.2	The Model Development	135
6.3	Fuzzy Data Processing	142
6.4	Functional Testing and Model Improvement	157

6.5 Performance Evaluation of the Model 157

6.6 Summarization 158

References 158

Glossary and Transcription 161

Appendix 163