

Theory and Applications of Natural Language Processing

Series Editors:

Graeme Hirst

Julia Hirschberg

Eduard Hovy

Mark Johnson

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

“Theory and Applications of Natural Language Processing” is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

- * Downloadable on your PC, e-reader or iPad
- * Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
- * Available online within an extensive network of academic and corporate R&D libraries worldwide
- * Never out of print thanks to innovative print-on-demand services
- * Competitively priced print editions for eBook customers thanks to MyCopy service <http://www.springer.com/librarians/e-content/mycopy>

More information about this series at
<http://www.springer.com/series/8899>

Carlos Ramisch

Multiword Expressions Acquisition

A Generic and Open Framework

 Springer

Carlos Ramisch
Aix Marseille University
Marseille
France

ISSN 2192-032X ISSN 2192-0338 (electronic)
ISBN 978-3-319-09206-5 ISBN 978-3-319-09207-2 (eBook)
DOI 10.1007/978-3-319-09207-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014950798

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved 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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my parents

Preface

The work described in this book was mostly carried out between 2009 and 2012 in Grenoble (France) and Porto Alegre (Brazil) as part of my PhD.¹ For their guidance and support, I would like to thank my advisers, Aline Villavicencio and Christian Boitet, who are true inspirations for me. I would also like to thank the colleagues of the Federal University of Rio Grande do Sul (UFRGS) and of the University of Grenoble. I thank Eric Wehrli and Gaël Dias for their comments and suggestions as *rapporteurs* of my thesis, as well as the other members of the jury: Yves Lepage, Helena de Medeiros Caseli, Rosa Vicari, Renata Vieira and Mathieu Mangeot. I am thankful to my Springer editor, Federica Corradi dell'Acqua, and the anonymous reviewers, who provided constructive feedback and helped improve my work.

This book is not as much a personal achievement as it is the result of a collective effort. Therefore, I am grateful to the people who contributed to it, in particular the co-authors of papers discussed in this book: Aline Villavicencio, Christian Boitet, Magali Sanches Duran, Evita Linardaki, Mathieu Mangeot, Cassia Trojahn dos Santos, Renata Vieira, Sandra Maria Aluísio, Maria José Finatto, Roger Granada, Marco Idiart, Lucelene Lopes, Helena de Medeiros Caseli, Laurent Besacier and Alexander Kobzar. I had interesting discussions and received valuable suggestions from Emmanuelle Esperança-Rodier, Paulo Schreiner, Rodrigo Wilkens, Valérie Bellynck, Sara Stymne, Lucia Specia, Agata Savary, Violeta Seretan, Dimitra Anastasiou, Preslav Nakov, Paul Cook, Kyo Kageura, Francesca Bonin and Muntsa Padró.

Many improvements in the `mwetoolkit` would not have been possible without the help of Sandra Castellanos, Maitê Dupont, Vitor De Araujo and Alexander Kobzar. I am specially indebted to Vitor for his programming skills and perspicacity. I would like to express my gratitude to all the anonymous users who downloaded the `mwetoolkit` and also to the non-anonymous ones who provided me with

¹Funding: French Ministry of Higher Education & Research, CAMELEON (CAPES-COFECUB 707-11), AIM-WEST (FAPERGS-INRIA 1706-2551/13-7), LIG, PPGC UFRGS, LICIA lab.

valuable feedback: Spence Green, Julien Corman, Agnès Tutin, Olivier Kraif, Cleci Bevilacqua, Anna Maciel, Guilherme Pilotti and Lis Kenashiro.

For their help as proofreaders, I thank my sister Renata and my partner Antoine. Along with the other members of my family and my friends, they have been extremely patient and supportive, particularly during the redaction of my thesis. This book is dedicated to my parents, who are unconditional fans, and especially to my father, who always said I should write a book some day.

Marseille, France
April 2014

Carlos Ramisch

Contents

1 Introduction	1
1.1 Motivations	1
1.1.1 What Are Multiword Expressions?	1
1.1.2 Why Do They Matter?	4
1.1.3 What Happens If We Ignore Them?	7
1.2 A New Framework for MWE Treatment	9
1.2.1 Hypotheses	9
1.2.2 Goals	10
1.2.3 Guiding Principles	11
1.3 Chapters Outline	14
1.4 Summary	16
References	17

Part I Multiword Expressions: A Tough Nut to Crack

2 Definitions and Characteristics	23
2.1 A Brief History	23
2.1.1 Theoretical Linguistics	24
2.1.2 Computational Linguistics	26
2.2 Defining MWEs	28
2.2.1 What Is a Word?	28
2.2.2 What Is a MWE?	29
2.2.3 A Note on Terminology	31
2.3 Characteristics and Characterisations	34
2.3.1 The Compositionality Continuum	34
2.3.2 Derived MWE Properties	36
2.3.3 Existing MWE Typologies	39
2.3.4 A Simplified Typology	41

2.4	A Snapshot of the Research Field	45
2.5	Summary	46
	References	47
3	State of the Art in MWE Processing	53
3.1	Elementary Notions	53
3.1.1	Linguistic Processing: Analysis	54
3.1.2	Word Frequency Distributions	57
3.1.3	<i>N</i> -Grams, Language Models and Suffix Arrays	60
3.1.4	Lexical Association Measures	63
3.2	Methods for Automatic MWE Acquisition	70
3.2.1	Monolingual Methods	71
3.2.2	Bi- and Multilingual Methods	74
3.2.3	Existing Tools	76
3.3	Other Tasks Related to MWE Processing	80
3.3.1	Interpretation	80
3.3.2	Disambiguation	83
3.3.3	Representation	84
3.3.4	Applications	86
3.4	Summary	91
	References	93

Part II MWE Acquisition

4	Evaluation of MWE Acquisition	105
4.1	Evaluation Context	106
4.1.1	Evaluation Axes	106
4.1.2	Evaluation Measures	109
4.1.3	Annotation	111
4.2	Acquisition Contexts	114
4.2.1	Characteristics of Target Constructions	115
4.2.2	Characteristics of Corpora	116
4.2.3	Existing Resources	119
4.3	Discussion	119
4.4	Summary	121
	References	122
5	A New Framework for MWE Acquisition	127
5.1	The <i>mwetoolkit</i> Framework	127
5.1.1	General Architecture	128
5.1.2	Modules	130
5.1.3	Discussion	138
5.2	A Toy Experiment	141
5.2.1	Candidate Extraction	141
5.2.2	Candidate Filtering	142
5.2.3	Results	144

- 5.3 Comparison with Related Approaches 145
 - 5.3.1 Related Approaches 145
 - 5.3.2 Comparison Setup 146
 - 5.3.3 Results..... 147
- 5.4 Summary 152
- References 154

Part III Applications

- 6 Application 1: Lexicography** 159
 - 6.1 A Dictionary of Nominal Compounds in Greek 159
 - 6.1.1 Greek Nominal Compounds..... 160
 - 6.1.2 Automatic Acquisition Setup 162
 - 6.1.3 Results..... 163
 - 6.2 A Dictionary of Complex Predicates in Portuguese 166
 - 6.2.1 Portuguese Complex Predicates..... 167
 - 6.2.2 Automatic Acquisition Setup 169
 - 6.2.3 Results..... 171
 - 6.3 Summary 176
 - References 178
- 7 Application 2: Machine Translation** 181
 - 7.1 A Brief Introduction to SMT..... 183
 - 7.2 Evaluation of Phrasal Verb Translation..... 186
 - 7.2.1 English Phrasal Verbs 187
 - 7.2.2 Translation Setup 189
 - 7.2.3 Results..... 192
 - 7.3 Summary 197
 - References 197
- 8 Conclusions** 201
 - References 204
- A Extended List of Translation Examples** 207
- B Resources Used in the Experiments** 209
 - B.1 Data 209
 - B.1.1 Monolingual Corpora 209
 - B.1.2 Multilingual Corpora 209
 - B.2 Software..... 210
 - B.2.1 Analysis Tools 210

C	The mwetoolkit: Documentation	211
C.1	Design Choices	211
C.2	Installing the mwetoolkit	212
C.2.1	Windows	212
C.2.2	Linux and Mac OS	212
C.2.3	Mac OS Dependencies	213
C.2.4	Testing Your Installation	213
C.3	Getting Started	213
C.3.1	An Example	214
C.4	Defining Patterns for Extraction	216
C.4.1	Literal Matches	216
C.4.2	Repetitions and Optional Elements	216
C.4.3	Ignoring Parts of the Match	217
C.4.4	Backpatterns	218
C.4.5	Syntactic Patterns	218
C.5	Preprocessing a Corpus Using TreeTagger	219
C.5.1	Installing TreeTagger	219
C.5.2	Converting TreeTagger's Output to XML	219
C.6	Preprocessing a Corpus Using RASP	220
C.6.1	Installing RASP	220
C.6.2	Converting RASP's Output to XML	220
C.7	Examples of XML Files	220
C.8	Developers	221
D	Tagsets for POS and Syntax	223
D.1	Generic POS Tagset	223
D.2	RASP English POS Tagset	223
D.3	RASP English Grammatical Relations	226
D.4	TreeTagger English POS Tagset	227
E	Detailed Lexicon Descriptions	229
E.1	Sentiment Verbs Extracted from Brazilian WordNet	229
E.2	Sentiment Nouns	230

Acronyms

AM	Association measure
CP	Complex predicate
DTD	Document type definition
GS	Gold standard
LM	Language model
LNRE	Large number of rare events
LSF	Lexico-semantic function
LVC	Light verb construction
MAP	Mean averaged precision
MLE	Maximum likelihood estimation
MTT	Meaning-text theory
MWE	Multiword expression
MWT	Multiword term
NLP	Natural language processing
POS	Part of speech
P	Precision
PV	Phrasal verb
R	Recall
SVC	Support verb construction
SVM	Support vector machine
TP	True positive
VPC	Verb-particle construction
XML	Extended markup language

Association Measures

dice	Dice's coefficient
ll	Log-likelihood ratio
mle	Maximum likelihood estimator
pmi	Pointwise mutual information
t-score	Student's t test statistic

Applications

IR	Information retrieval
MT	Machine translation
OCR	Optical character recognition
PB-SMT	Phrase-based statistical machine translation
SMT	Statistical machine translation
SRL	Semantic role labelling
WSD	Word sense disambiguation

Conferences

ACL	Annual Meeting of the Association for Computational Linguistics
COLING	International Conference on Computational Linguistics
EACL	European Chapter of the Association for Computational Linguistics
LREC	Language Resources and Evaluation Conference
NAACL	North American Chapter of the Association for Computational Linguistics

Corpora

BNC	British national corpus
EP	Europarl corpus
PLN-BR	Corpus of the project Processamento de Linguagem Natural—Brasil

Language Codes

e1	Greek
en	English
fr	French
pt	Portuguese
pt - BR	Brazilian Portuguese

Symbols

*	Ungrammatical construction
?	Unnatural construction