# Theory and Applications of Natural Language Processing

## 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



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

[1] Funding: 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 Carlos Ramisch
April 2014

# Contents

# 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**

| | |
|---|---|
| `el` | Greek |
| `en` | English |
| `fr` | French |
| `pt` | Portuguese |
| `pt-BR` | Brazilian Portuguese |

**Symbols**

| | |
|---|---|
| * | Ungrammatical construction |
| ? | Unnatural construction |