Lecture Notes in Artificial Intelligence 7243

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

LNAI Founding Series Editor

Helena Caseli Aline Villavicencio
António Teixeira Fernando Perdigão (Eds.)

Computational Processing of the Portuguese Language

10th International Conference, PROPOR 2012
Coimbra, Portugal, April 17-20, 2012
Proceedings

Springer

Volume Editors

Helena Caseli
UFSCAR, São Carlos, SP, Brazil
E-mail: helenacaseli@dc.ufscar.br

Aline Villavicencio
UFRGS, Porto Alegre, RS, Brazil
E-mail: alinev@gmail.com

António Teixeira
Universidade de Aveiro, Aveiro, Portugal
E-mail: ajst@ua.pt

Fernando Perdigão
Universidade de Coimbra, Coimbra, Portugal
E-mail: fp@deec.uc.pt

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-28884-5 e-ISBN 978-3-642-28885-2
DOI 10.1007/978-3-642-28885-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012933122

CR Subject Classification (1998): I.2, H.3, H.4, I.4, I.5, H.2.8

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

The International Conference on Computational Processing of Portuguese – PROPOR – is the main event in the area of natural language processing that is focused on Portuguese and the theoretical and technological issues related to this language. It welcomes contributions for both written and spoken language processing.

The event is hosted in Brazil and in Portugal. The meetings have been held in Lisbon/Portugal (1993), Curitiba/Brazil (1996), Porto Alegre/Brazil (1998), Évora/Portugal (1999), Atibaia/Brazil (2000), Faro/Portugal (2003), Itatiaia/Brazil (2006), Aveiro/Portugal (2008), and Porto Alegre/Brazil (2010).

This meeting has been a highly productive forum for the progress of this area and to foster the cooperation among the researchers working on the automated processing of the Portuguese language. PROPOR brings together research groups promoting the development of methodologies, resources, and projects that can be shared among all researchers and practitioners in the field.

The tenth edition of this event was held at the University of Coimbra, Coimbra, Portugal. It had two main tracks: one for language processing and another for speech processing. This event hosted a special Demonstration Session and a satellite event named "Págico," consisting in an evaluation contest for non-trivial information seeking in Portuguese using Wikipedia as target. This edition of PROPOR featured two invited talks by internationally renowned researchers as well as tutorials on symbolic and statistical approaches to natural language processing and analysis and visual feedback of the singing voice.

A total of 86 submissions were received, 61 for the language track and 25 for the speech track, by authors in worldwide institutions from countries like Brazil, China, Germany, Portugal, and Spain. Each submission was evaluated by at least three members from a multidisciplinary and international scientific committee.

This volume gathers a selection of the 47 best papers accepted to be presented at this meeting, of which 24 are full papers, corresponding to an acceptance rate of 27%. These papers cover the areas related to automatic acquisition of information, linguistic description and processing, language resources, language applications and speech production, speech processing and applications.

We would like to express our thanks to everyone involved in the organization of the event, to the scientific committee members for their excellent work, to the researchers who kindly accepted to contribute to the event by delivering tutorials and invited talks, and to the institutions, organizations, and funding

agencies which allowed the realization of this event, namely, University of Coimbra, IT (Instituto de Telecomunicações), FCT (Portuguese National Founding Agency), ISCA (International Speech Communication Association), SIG-IL (the ISCA Special Group on Iberian Languages), CEPLN (the Special Interest Group on Natural Language Processing of the Brazilian Computer Society), and ACL (Association for Computational Linguistics).

April 2012

Helena de Medeiros Caseli
Aline Villavicencio
António Teixeira
Fernando Perdigão

Organization

General Chair

Fernando Perdigão	Universidade de Coimbra/IT, Portugal

Program Chairs

Aline Villavicencio	UFRGS, Brazil – Language
António Teixeira	Universidade de Aveiro/IEETA, Portugal – Speech

Editorial Chair

Helena de Medeiros Caseli	UFSCAR, Brazil

Demo Session Chair

Alberto Abad	L2F INESC-ID, Portugal

PhD and MSc/MA Dissertation Contest Chair

Jorge Baptista	Universidade do Algarve, Portugal

Local Organizing Committee

Luís Sá	UC/IT, Portugal
Sara Candeias	IT, Portugal
Ana R. Luís	UC/CELGA, Portugal
Carla Lopes	IPLEI/IT, Portugal
Hugo Gonçalo Oliveira	UC/CISUC, Portugal

Program Committee

Alberto Abad	INESC-ID, Portugal
Alberto Simões	UM, Portugal
Aldebaro Klautau	UFPA, Brazil
Alexandre Agustini	PUC-RS, Brazil
Aline Villavicencio	UFRGS, Brazil
Amália Andrade	UL, Portugal
Amália Mendes	UL, Portugal

Ana Bocorny	PUC-RS, Brazil
Ana Luís	UC, Portugal
Andre Adami	UCS, Brazil
Andreia Rauber	UCPEL, Brazil
Antonio Bonafonte	UPC, Spain
António Branco	UL, Portugal
Antonio Rubio	UG, Spain
António Serralheiro	INESC-ID, Portugal
António Teixeira	Universidade de Aveiro, Portugal
Ariadne Carvalho	Unicamp, Brazil
Ariani Di Felippo	UFSCAR, Brazil
Belinda Maia	UP, Portugal
Bento da Silva	UNESP, Brazil
Berthold Crysmann	CNRS Paris-Diderot, France
Carlos Prolo	PUC-RS, Brazil
Carlos Teixeira	UL, Portugal
Carmen García Mateo	UV, Spain
Caroline Gasperin	TouchType, UK
Caroline Hagège	Xerox Research Centre, France
Catarina Oliveira	Universidade de Aveiro, Portugal
Ciro Martins	Universidade de Aveiro, Portugal
Cristiane Killian	UFRGS, Brazil
Daniela Braga	Microsoft, China
Dante Barone	UFRGS, Brazil
Diana Santos	University of Oslo, Norway
Doroteo Torre Toledano	UAM, Spain
Eduardo Lleida	UZ, Spain
Eric Laporte	Université Paris Est, France
Eva Navas	UBC, Spain
Fabio Kepler	USP, Brazil
Fábio Violaro	Unicamp, Brazil
Fernando Resende	UFRJ, Brazil
Gaël Harry Dias	UBI, Portugal
Gladis Almeida	UFSCAR, Brazil
Helena de Medeiros Caseli	UFSCAR, Brazil
Irene Rodrigues	UE, Portugal
Isabel Falé	Universidade Aberta, Portugal
Isabel Trancoso	INESC-ID/IST, Portugal
Ivandré Paraboni	USP, Brazil
Jean-Luc Minel	Université de Paris X, France
João Balsa	UL, Portugal
João Luís Rosa	USP-SC, Brazil
João Paulo Neto	INESC-ID/IST, Portugal
João Veloso	UP, Portugal
Joaquim Ferreira da Silva	UNL, Portugal
Joaquim Llisterri	UAB, Spain

Jorge Baptista	Universidade do Algarve, Portugal
José Gabriel Lopes	UNL, Portugal
José João Almeida	UM, Portugal
Julia Hirschberg	Columbia University, USA
Laura Alonso Alemany	University National of Cordoba, Argentina
Leandro Oliveira	Embrapa, Brazil
Leandro Wives	UFRGS, Brazil
Lúcia Rino	UFSCAR, Brazil
Lucia Specia	University of Wolverhampton, UK
Luís Oliveira	INESC-ID, Portugal
Luís Sá	UC, Portugal
Luísa Coheur	INESC-ID/IST, Portugal
Luiz Pizzato	University of Sydney, Australia
Magali Duran	USP-SC, Brazil
Mara Abel	UFRGS, Brazil
Marcelo Finger	USP, Brazil
Marco Gonzalez	PUC-RS, Brazil
Maria das Graças Volpe Nunes	USP-SC, Brazil
Maria Helena Mira Mateus	ILTEC, Portugal
Maria José Finatto	UFRGS, Brazil
Mário Silva	INESC-ID/IST, Portugal
Michel Gagnon	Ecole Polytechnique, Canada
Miguel Sales Dias	Microsoft-MLDC, Portugal
Nuno Cavalheiro Marques	UNL, Portugal
Nuno Mamede	INESC-ID/IST, Portugal
Pablo Gamallo	University of Santiago de Compostela, Spain
Palmira Marrafa	UL, Portugal
Paulo Gomes	UC, Portugal
Paulo Quaresma	UE, Portugal
Plínio Barbosa	Unicamp, Brazil
Ranniery Maia	Toshiba, UK
Renata Vieira	PUC-RS, Brazil
Ricardo Ribeiro	INESC-ID/ISCTE-IUL, Portugal
Robert Dale	Macquarie University, Australia
Ronaldo Martins	Univas, Brazil
Rove Chishman	Unisinos, Brazil
Rubén San-Segundo	UPM, Spain
Ruy Luiz Milidiú	PUC-Rio, Brazil
Sandra Aluisio	USP-SC, Brazil
Sergio Freitas	UnB, Brazil
Solange Rezende	USP-SC, Brazil
Stanley Loh	UCPEL, Brazil
Steven Bird	University of Melbourne, Australia
Thiago Pardo	USP-SC, Brazil
Tracy Holloway King	Microsoft, USA

Valéria Feltrim	UEM, Brazil
Vera Strube de Lima	PUC-RS, Brazil
Violeta Quental	PUC-Rio, Brazil
Vitor Rocio	Universidade Aberta, Portugal
Viviane Moreira	UFRGS, Brazil

Steering Committee

António Teixeira	Universidade de Aveiro, Portugal (Chair)
Fernando Perdigão	Universidade de Coimbra, Portugal
Jorge Baptista	Universidade do Algarve, Portugal
Renata Vieira	PUC-RS, Brazil
Violeta Quental	PUC-RJ, Brazil

Table of Contents

Phonology, Morphology and POS-Tagging

Verb Analysis in a Highly Inflective Language with an MFF Algorithm 1
António Branco and Filipe Nunes

Automatic Analysis of Portuguese Verb Morphology: Solving Ambiguities Caused by Thematic Vowel Allomorphs 12
Vera Vasilévski, Leonor Scliar-Cabral, and Márcio José Araújo

Coordination of *-mente* Ending Adverbs in Portuguese: An Integrated Solution 24
Jorge Baptista, Lucas Nunes Vieira, Cláudio Diniz, and Nuno Mamede

Morphosyntactic Analysis of Language in Children with Autism Spectrum Disorder 35
Raquel Reis and António Teixeira

Lince, an End User Tool for the Implementation of the Spelling Reform of Portuguese 46
José Pedro Ferreira, António Lourinho, and Margarita Correia

Searching a Mixed Corpus in the Light of the New Portuguese Orthographic Norm 56
Gracinda Carvalho, Isabel Falé, David Martins de Matos, and Vitor Rocio

Acquisition

Extraction of Bilingual Cognates from Wikipedia 63
Pablo Gamallo and Marcos Garcia

Corpus-Based Acquisition of Support Verb Constructions for Portuguese 73
Britta D. Zeller and Sebastian Padó

Improving Portuguese Term Extraction 85
Lucelene Lopes and Renata Vieira

A Method for Automatically Extracting Domain Semantic Networks from Wikipedia 93
Clarissa Castellã Xavier and Vera Lúcia Strube de Lima

Extracting Temporal Information from Portuguese Texts 99
Francisco Costa and António Branco

It Is the Time for Portuguese Texts! 106
Olga Craveiro, Joaquim Macedo, and Henrique Madeira

Language Resources

A Large Portuguese Corpus On-Line: Cleaning and Preprocessing 113
Michel Généreux, Iris Hendrickx, and Amália Mendes

Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing 121
Alberto Simões, Álvaro Iriarte Sanromán, and José João Almeida

Towards a Common Sense Base in Portuguese for the Linked Open Data Cloud 128
Vládia Pinheiro, Vasco Furtado, Tarcisio Pequeno, and Caio Ferreira

Linguistic Description, Syntax and Parsing

Weak Object Pronouns in Brazilian Portuguese: An LFG Analysis 139
Ana R. Luís

Entropy-Guided Feature Generation for Structured Learning of Portuguese Dependency Parsing 146
Eraldo R. Fernandes and Ruy L. Milidiú

Bayesian Induction of Syntactic Language Models for Brazilian Portuguese 157
Daniel Emilio Beck and Helena de Medeiros Caseli

Automatic Generation of *Cloze* Question Stems 168
Rui Correia, Jorge Baptista, Maxine Eskenazi, and Nuno Mamede

Semantics

Toponym Disambiguation Using Ontology-Based Semantic Similarity . . . 179
David S. Batista, João D. Ferreira, Francisco M. Couto, and Mário J. Silva

Automatic Hyponymy Identification from Brazilian Portuguese Texts . . . 186
Leonardo Sameshima Taba and Helena de Medeiros Caseli

Semantic Role Labeling for Portuguese – A Preliminary Approach – 193
João Sequeira, Teresa Gonçalves, and Paulo Quaresma

An Architecture for Semantic Role Labeling on Portuguese 204
Erick Rocha Fonseca and João Luís G. Rosa

Towards Semi-supervised Brazilian Portuguese Semantic Role Labeling: Building a Benchmark 210
Fernando Emilio Alva-Manchego and João Luís G. Rosa

Opinion Analysis

Building a Sentiment Lexicon for Social Judgement Mining 218
Mário J. Silva, Paula Carvalho, and Luís Sarmento

A Bootstrapping Algorithm for Learning the Polarity of Words 229
António Paulo Santos, Hugo Gonçalo Oliveira, Carlos Ramos, and Nuno C. Marques

The Role of Language Registers in Polarity Propagation 235
António Paulo Santos, Hugo Gonçalo Oliveira, Carlos Ramos, and Nuno C. Marques

Sentiment Analysis on Twitter Data for Portuguese Language 241
Marlo Souza and Renata Vieira

Natural Language Processing Applications

REAP.PT Serious Games for Learning Portuguese 248
André Silva, Cristiano Marques, Jorge Baptista, Alfredo Ferreira Jr., and Nuno Mamede

Graph-Based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information 260
Rafael Ribaldo, Ademar Takeo Akabane, Lucia Helena Machado Rino, and Thiago Alexandre Salgueiro Pardo

Clustering and Categorization of Brazilian Portuguese Legal Documents 272
Luis Otávio de Colla Furquim and Vera Lúcia Strube de Lima

SIGA, a System to Manage Information Retrieval Evaluations 284
Luis Costa, Cristina Mota, and Diana Santos

E-commerce Market Analysis from a Graph-Based Product Classifier . . . 291
Andréa Britto Mattos, Marcelo Van Kampen, Camila Carriço, André Ricardo Dias, and Alexandre Crivellaro

A Description Logic for InferenceNet.Br 298
Wellington Franco, Thiago Alves, Henrique Viana, and João Alcântara

Speech Production and Phonetics

Real-Time MRI for Portuguese Database, Methods and Applications ... 306
António Teixeira, Paula Martins, Catarina Oliveira, Carlos Ferreira, Augusto Silva, and Ryan Shosted

Production and Modeling of the European Portuguese Palatal Lateral ... 318
António Teixeira, Paula Martins, Catarina Oliveira, and Augusto Silva

A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches ... 329
Plínio A. Barbosa and Wellington da Silva

Constructing Physically Realistic VCV Stimuli for the Perception of Stop Voicing in European Portuguese ... 338
Daniel Pape, Luis M.T. Jesus, and Pascal Perrier

Automatic Phonetic Transcription by Phonological Derivation ... 350
Marcos Garcia and Isaac J. González

Speech Resources

The C-ORAL-BRASIL I: Reference Corpus for Informal Spoken Brazilian Portuguese ... 362
Tommaso Raso and Heliana Mello

A European Portuguese Children Speech Database for Computer Aided Speech Therapy ... 368
Carla Lopes, Arlindo Veiga, and Fernando Perdigão

Baseline Acoustic Models for Brazilian Portuguese Using CMU Sphinx Tools ... 375
Rafael Oliveira, Pedro Batista, Nelson Neto, and Aldebaro Klautau

Speech Processing and Applications

A Fishervoice-SVM Language Identification System ... 381
Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Summarizing Speech by Contextual Reinforcement of Important Passages ... 392
Ricardo Ribeiro and David Martins de Matos

Incorporating ASR Information in Spoken Dialog System Confidence Score ... 403
José Lopes, Maxine Eskenazi, and Isabel Trancoso

Transcription of Multi-variety Portuguese Media Contents 409
Alberto Abad, Hugo Meinedo, Isabel Trancoso, and João Neto

Towards Automatic Classification of Speech Styles 421
Arlindo Veiga, Sara Candeias, Dirce Celorico, Jorge Proença, and Fernando Perdigão

Author Index 427