Proceedings of the 11th International Conference on Computational Processing of the Portuguese Language - PROPOR

http://www.producao.usp.br/handle/BDPI/48613

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo
Lecture Notes in Artificial Intelligence 8775

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel
   University of Alberta, Edmonton, Canada

Yuzuru Tanaka
   Hokkaido University, Sapporo, Japan

Wolfgang Wahlster
   DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann
   DFKI and Saarland University, Saarbrücken, Germany
Preface

The International Conference on Computational Processing of Portuguese (PRO- POR), in its 2014 edition, continued being the most important natural language processing scientific event dedicated to the Portuguese language and its theoretical and technological advances, simultaneously dealing both with its spoken and written dimensions. This event is hosted every two years, alternating between Brazil and Portugal. Previous events were 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), Porto Alegre/Brazil (2010), and Coimbra/Portugal (2012). The conference brings together researchers and practitioners in the field, sharing their expertise, divulging their results, promoting methodologies and exchanging resources, forming a very active and vibrant community, internationally recognized for their excellence of standards and acutely aware of their role in the growing impact of Portuguese in international fora. This 11th edition of PROPOR took place in São Carlos/SP, in the south-east of Brazil, under the aegis of the Institute of Mathematical and Computer Sciences (ICMC) of the University of São Paulo (USP) and the Interinstitutional Center for Computational Linguistics (NILC). The event also featured the third edition of the MSc/MA and PhD Dissertation Contest, which selects the best new academic research in Portuguese NLP, as well as a workshop for demonstration of software and resources for Portuguese processing.

Three keynote speakers honored the event with their lectures: Dr. Advaith Siddharthan (University of Aberdeen), Dr. Andreas Stolcke (Microsoft Research), and Dr. Michael Riley (Google Inc.).

A total of 63 submissions were received for the main event, totaling 91 authors from many institutions worldwide, such as Algeria, Brazil, France, Norway, Portugal, The Netherlands, and USA.

This volume brings together a selection of the 35 best papers accepted at this meeting: 14 full papers and 19 short papers. The acceptance rate was 22%. To these, the two best papers from the Phd and MSc/MA dissertations contest were added.

In this volume, the papers are organized thematically and include the most recent developments in speech language processing and applications, linguistic description, syntax and parsing, ontologies, semantics and lexicography, corpora and language resources, natural language processing, tools and applications.

Our sincere thanks to every person and institution involved in the complex organization of this event, especially the members of the scientific committee of the main event, the dissertations contest and the associated workshops, the invited speakers, and the general organization staff.
We are also grateful to the agencies and organizations that supported and promoted the event, namely, the Brazilian Computer Society (SBC) and its Special Interest Group on NLP (CEPLN), the São Paulo Research Foundation (FAPESP), the Coordination for the Improvement of Higher Level Personnel (CAPES), the International Speech Communication Association (ISCA), and Samsung Eletrônica da Amazônia, Ltda.

October 2014

Jorge Baptista
Nuno Mamede
Sara Candeias
Ivandré Paraboni
Thiago A.S. Pardo
Maria das Graças Volpe Nunes
Organization

General Chairs
Thiago Alexandre
    Salgueiro Pardo USP-ICMC, Brazil
Maria das Graças Volpe Nunes USP-ICMC, Brazil

Technical Program Chairs
Sara Candeias IT-University of Coimbra, Portugal
Ivandré Paraboni USP-EACH, Brazil

Editorial Chairs
Jorge Baptista University of Algarve, Portugal
Nuno Mamede IST/INESC-ID, Portugal

Workshop Chair
Vera Lúcia Strube de Lima PUCRS, Brazil

Demos Chairs
Arnaldo Candido Jr. UFSCar, Brazil
Fábio Natanael Kepler UNIPAMPA, Brazil

PhD and MSc/MA Dissertation Contest Chairs
Helena de Medeiros Caseli UFSCar-DC, Brazil
David Martins de Matos IST/INESC-ID, Portugal

Local Organizing Committee
Sandra Aluíso USP-ICMC, Brazil
Lúcia Helena Machado Rino UFSCar-DC, Brazil
Magali Sanches Duran USP-ICMC, Brazil
Pedro Paulo Balage Filho USP-ICMC, Brazil
Program Committee

Albert Gatt University of Malta, Malta
Alberto Abad INESC-ID, Portugal
Alberto Simões Universidade do Minho, Portugal
Alexandre Agustini PUCRS, Brazil
Alíne Villavicencio UFRGGS, Brazil
Amália Andrade Univ. de Lisboa, Portugal
Amália Mendes Universidade de Lisboa, Portugal
Ana Luís Universidade de Coimbra, Portugal
Anabela Barreiro INESC-ID, Portugal
Andreia Bonfante UFMT, Brazil
Andreia Rauber Appen, USA
António Branco Universidade de Lisboa, Portugal
António Joaquim Serralheiro AM/INESC-ID, Portugal
António Teixeira Universidade de Aveiro, Portugal
Ariadne Carvalho UNICAMP, Brazil
Ariani Di Felippo UFSCAR, Brazil
Bento da Silva UNESP, Brazil
Berthold Crysmann CNRS, France
Brett Drury USP-SC, Brazil
Carla Alexandra Calado Lopes IT/IPL, Portugal
Carlos Prolo UFRN, Brazil
Caroline Gasperin TouchType, UK
Daniela Braga VoiceBox, USA
Diana Santos University of Oslo, Norway
Doroteo Torre Toledano UAM, Spain
Eduardo Lleida UZ, Spain
Eric Laporte Université Paris Est, France
Fábio Natanael Kepler UNIPAMPA, Brazil
Fernando Batista INESC-ID/ISCTE-IUL, Portugal
Fernando Perdigão IT/Universidade de Coimbra, Portugal
Fernando Resende UFRJ, Brazil
Gaël Harry Dias Université de Caen, Basse-Normandie, France
Gladis Almeida UFSCAR, Brazil
Helena de Medeiros Caseli UFSCAR, Brazil
Helena Moniz L2F/INESC-ID, Portugal
Hugo Meinedo INESC-ID, Portugal
Irina Temnikova QCRI, Qatar
Isabel Falé Universidade Aberta, Portugal
Isabel Trancoso IST/INESC-ID, Portugal
João Balsa Universidade de Lisboa, Portugal
João Luís Rosa USP-ICMC, Brazil
João Paulo Neto IST/INESC-ID, Portugal
Joaquim Llisterri UAB, Spain
Jorge Baptista Universidade do Algarve, Portugal
José João Almeida Universidade do Minho, Portugal
Kees van Deemter University of Aberdeen, UK
Laura Alonso Alemany Universidad Nacional de Córdoba, Argentina
Leandro Oliveira Embrapa, Brazil
Luciana Benotti Universidad Nacional de Córdoba, Argentina
Luis A. Pineda UNAM, Mexico
Luiz Pizzato University of Sydney, Australia
Magali Sanches Duran USP-SC, Brazil
Marcelo Finger USP, Brazil
Maria das Graças Volpe Nunes USP-SC, Brazil
Maria José Finatto UFRGS, Brazil
Mário J. Silva IST/INESC-ID, Portugal
Michel Gagnon Ecole Polytechnique, Canada
Norton Trevisan Roman USP-EACH, Brazil
Nuno Cavalheiro Marques UNL, Portugal
Nuno Mamede IST/INESC-ID, Portugal
Palmira Marrafa Universidade de Lisboa, Portugal
Paulo Gomes Universidade de Coimbra, Portugal
Paulo Quresma Universidade de Évora, Portugal
Plínio Barbosa Unicamp, Brazil
Renata Vieira PUCRS, Brazil
Ricardo Ribeiro INESC-ID/ISCTE-IUL, Portugal
Ronaldo Martins Univas, Brazil
Rubén San-Segundo UPM, Spain
Ruy Luiz Milidiú PUC-Rio, Brazil
Sandra Aluísio USP-SC, Brazil
Sara Candeias IT/Universidade de Coimbra, Portugal
Solange Rezende USP-SC, Brazil
Steven Bird University of Melbourne, Australia
Ted Pederson University of Minnesota, USA
Thiago A.S. Pardo USP-SC, Brazil
Thomas Pellegrini Université de Toulouse III-Paul Sabatier, France
Valéria Feltrim UEM, Brazil
Vera Lúcia Strube de Lima PUCRS, Brazil
Violeta Quental PUC-Rio, Brazil
Vitor Rocio Universidade Aberta, Portugal
Wilker Aziz University of Sheffield, UK
Additional Reviewers

David Batista  
Denise Hogetop  
Derek Wong  
Gracinda Carvalho  
João Filgueiras  
Marcelo Criscuolo  
Mário Rodrigues  
Raquel Amaro  
Sara Mendes

INESC-ID, Portugal  
PUCRS, Brazil  
University of Macau, China  
Universidade Aberta, Portugal  
INESC-ID, Portugal  
USP, Brazil  
Universidade de Aveiro, Portugal  
Universidade de Lisboa, Portugal

Steering Committee

Jorge Baptista  
Cláudia Freitas  
Fernando Perdigão  
Renata Vieira  
Thiago Alexandre  
Salgueiro Pardo

Universidade do Algarve, Portugal (Chair)  
PUC-Rio/Linguateca, Brazil  
IT/Universidade de Coimbra, Portugal  
PUCRS, Brazil  
USP-ICMC, Brazil
# Table of Contents

## Speech Language Processing and Applications

Automatically Recognising European Portuguese Children’s Speech: Pronunciation Patterns Revealed by an Analysis of ASR Errors ........ 1  
*Annika Hämäläinen, Hyongsil Cho, Sara Candeias, Thomas Pellegrini, Alberto Abad, Michael Tjalve, Isabel Trancoso, and Miguel Sales Dias*

Improving Speech Recognition through Automatic Selection of Age Group – Specific Acoustic Models ................................. 12  
*Annika Hämäläinen, Hugo Meinedo, Michael Tjalve, Thomas Pellegrini, Isabel Trancoso, and Miguel Sales Dias*

Characterizing Parkinson’s Disease Speech by Acoustic and Phonetic Features ................................................................. 24  
*Jorge Proença, Arlindo Veiga, Sara Candeias, João Lemos, Cristina Januário, and Fernando Perdigão*

Rule-Based Algorithms for Automatic Pronunciation of Portuguese Verbal Inflections .............................................................. 36  
*Vanessa Marquiafável, Christopher Shulby, Arlindo Veiga, Jorge Proença, Sara Candeias, and Fernando Perdigão*

Acoustic Similarity Scores for Keyword Spotting ........................ 48  
*Arlindo Veiga, Carla Lopes, Luís Sá, and Fernando Perdigão*

## Linguistic Description, Syntax and Parsing

JMorpher: A Finite-State Morphological Parser in Java for Android .... 59  
*Leonel F. de Alencar, Mardonio J.C. França, Katiuscia M. Andrade, Philipp B. Costa, Henrique S. Vasconcelos, and Francinaldo P. Madeira*

Tagging and Labelling Portuguese Modal Verbs ................................. 70  
*Paulo Quaresma, Amália Mendes, Iris Hendrickx, and Teresa Gonçalves*

Training State-of-the-Art Portuguese POS Taggers without Handcrafted Features ................................................................. 82  
*Cícero Nogueira dos Santos and Bianca Zadrozny*
General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese ...................................................... 94
Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo

Semi-supervised Parsing of Portuguese ...................................................... 102
Pablo Botton da Costa and Fabio Natanael Kepler

What We Can Learn from Looking at Profanity ........................................ 108
Gustavo Laboreiro and Eugénio Oliveira

Ontologies, Semantics and Lexicography

Extending a Lexicon of Portuguese Nominalizations with Data from Corpora .............................................................. 114
Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne Silva

Body-Part Nouns and Whole-Part Relations in Portuguese ...................... 125
Ilia Markov, Nuno Mamede, and Jorge Baptista

Proverb Variation: Experiments on Automatic Detection in Brazilian Portuguese Texts ........................................................................ 137
Amanda Rassi, Jorge Baptista, and Oto Vale

Using Cross-Linguistic Knowledge to Build VerbNet-Style Lexicons: Results for a (Brazilian) Portuguese VerbNet ..................................................... 149
Carolina Scarton, Magali Sanches Duran, and Sandra Maria Aluísio

The Creation of Onto.PT: A Wordnet-Like Lexical Ontology for Portuguese ...................................................................................... 161
Hugo Gonçalo Oliveira

Comparing Semantic Relatedness between Word Pairs in Portuguese Using Wikipedia ................................................................................ 170
Roger Granada, Cassia Trojahn, and Renata Vieira

On the Utility of Portuguese Term-Based Lexical-Semantic Networks ... 176
Hugo Gonçalo Oliveira

Semantic Textual Similarity of Portuguese-Language Texts: An Approach Based on the Semantic Inferentialism Model ........................................ 183
Vládia Pinheiro, Vasco Furtado, and Adriano Albuquerque

Temporal Aspects of Content Recommendation on a Microblog Corpus ................................................................................ 189
Caio Ramos Casimiro and Ivandré Paraboni
Development of a Lexical Resource Annotated with Semantic Roles for Portuguese ................................................................. 195
Leonardo Zilio

Corpora and Language Resources

brWaC: A WaCky Corpus for Brazilian Portuguese ......................... 201
Rodrigo Boos, Kassius Prestes, Aline Villavicencio, and Muntsa Padró

DeepBankPT and Companion Portuguese Treebanks in a Multilingual Collection of Treebanks Aligned with the Penn Treebank .............. 207
António Branco, Catarina Carvalheiro, Francisco Costa, Sérgio Castro, João Silva, Cláudia Martins, and Joana Ramos

Gramateca: Corpus-Based Grammar of Portuguese ....................... 214
Diana Santos

Natural Language Processing, Tools and Applications

Automatic Alignment of News Texts and Their Multi-document Summaries: Comparison among Methods ........................................ 220
Verónica Agostini, Roque Enrique López Condori, and Thiago Alexandre Salgueiro Pardo

Using Rhetorical Structure Theory and Entity Grids to Automatically Evaluate Local Coherence in Texts ............................... 232
Márcio de S. Dias, Valéria D. Feltrim, and Thiago Alexandre Salgueiro Pardo

Comparing NERP-CRF with Publicly Available Portuguese Named Entities Recognition Tools .................................................. 244
Daniela O.F. do Amaral, Evandro Fonseca, Lucelene Lopes, and Renata Vieira

Integrating Verbal Idioms into an NLP System ................................ 250
Jorge Baptista, Nuno Mamede, and Ilia Markov

Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology ...................................... 256
António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz

Words Temporality for Improving Query Expansion .......................... 262
Olga Craveiro, Joaquim Macedo, and Henrique Madeira
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Open Source Tool for Crowd-Sourcing the Manual Annotation of</td>
<td>268</td>
</tr>
<tr>
<td>Texts</td>
<td></td>
</tr>
<tr>
<td>Brett Drury, Paula C.F. Cardoso, Jorge Valverde-Rebaza,</td>
<td></td>
</tr>
<tr>
<td>Alan Valejo, Fabio Pereira, and Alneu de Andrade Lopes</td>
<td></td>
</tr>
<tr>
<td>Identification of Brazilian Portuguese Causative Verbs through a</td>
<td>274</td>
</tr>
<tr>
<td>Weighted Graph Classification Strategy</td>
<td></td>
</tr>
<tr>
<td>Brett Drury, Rafael Geraldeli Rossi, and Alneu de Andrade Lopes</td>
<td></td>
</tr>
<tr>
<td>Lausanne: A Framework for Collaborative Online NLP Experiments</td>
<td>280</td>
</tr>
<tr>
<td>Douglas Iacovelli, Michelle Reis Galindo, and Ivandré Paraboni</td>
<td></td>
</tr>
<tr>
<td>Alignment-Based Sentence Position Policy in a News Corpus for</td>
<td>286</td>
</tr>
<tr>
<td>Multi-document Summarization</td>
<td></td>
</tr>
<tr>
<td>Fernando Antônio Asevedo Nóbrega, Verônica Agostini,</td>
<td></td>
</tr>
<tr>
<td>Renata T. Camargo, Ariani Di Felippo, and</td>
<td></td>
</tr>
<tr>
<td>Thiago Alexandre Salgueiro Pardo</td>
<td></td>
</tr>
<tr>
<td>Identification of Related Brazilian Portuguese Verb Groups Using</td>
<td>292</td>
</tr>
<tr>
<td>Overlapping Community Detection</td>
<td></td>
</tr>
<tr>
<td>Alan Valejo, Brett Drury, Jorge Valverde-Rebaza, and</td>
<td></td>
</tr>
<tr>
<td>Alneu de Andrade Lopes</td>
<td></td>
</tr>
<tr>
<td><strong>Author Index</strong></td>
<td>299</td>
</tr>
</tbody>
</table>