

Luís Carlos dos Santos Marujo

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
lmajujo@cs.cmu.edu
luis.majujo@inesc-id.pt

Education Carnegie Mellon University, Pittsburgh, United States
 Instituto Superior Técnico, Lisbon, Portugal
Dual PhD candidate in Language and Information Technologies,
School of Computer Science (Carnegie Mellon Portugal program)
Advisors: Anatole Gershman, Jaime Carbonell, João P. Neto, and
David Matos

Instituto Superior Técnico, Lisbon, Portugal
C.A.S. in Information Systems and Computer Engineering,
Computer Science and Software Engineering Department
Overall: 18/20 (A) , July 2011

Instituto Superior Técnico, Lisbon, Portugal
MSc. Information Systems and Computer Engineering, Computer
Science and Software Engineering Department
Major: Software Engineering
Minor: Artificial Intelligence
Overall: 17/20 (Very Good), *Cum Laude*, July 2009
Thesis: 19/20 name: REAP.PT
Advisors: Nuno Mamede and Isabel Trancoso

Instituto Superior Técnico, Lisbon, Portugal
BSc. Information Systems and Computer Engineering, Computer
Science and Software Engineering Department
Overall: 16/20 (Very Good), July 2007

Summer School

LxMLS 2011- 1st Lisbon **Machine Learning** School at IST

Topic: Learning for the Web

The school covered a range of machine learning (ML) Topics, from theory to practice, that are important in solving natural language processing (NLP) problems that arise in the analysis and use of Web data.

S3MR – 2nd Summer School on **Social Media Retrieval**, Antalya, Turkey 2011

Topics:

- Content distribution over social/peer-to-peer networks
- Multimedia content analysis
- Automatic multimedia annotation/tagging
- Multimedia indexing/search/retrieval
- Implicit media tagging
- Social data analysis
- Collaborative tagging

Best Poster Award (Key-Phrase Extraction on Broadcast News)

Research Experience

Carnegie Mellon University,

August 2011 – Present

Graduate Research Assistant

Advisors: Anatole Gershman, Jaime Carbonell, and João P. Neto

Projects: euTV (European Project) and IdentityTracking (USA

NMEC project)

- Extending Supervised Key Phrase Extraction with pre-processing steps (light filtering and co-reference normalization) and semantic features.

- etc.

Spoken Language System Laboratory of INESC-ID Lisbon,
September 2010 – July 2011

Graduate Research Assistant

Advisor: Anatole Gershman, Jaime Carbonell, and João P. Neto

Programming Languages: **Java**, C++, Bash Scripting, Shell Scripting,

Projects: euTV (European Project) and PT-STAR (National Portuguese Project):

- Developing a prediction method to popularity of News Stories: number of clicks during an hour in Sapo portal (Most visited Portuguese Web Portal – www.sapo.pt) - 4th out of 26th systems in the 1st Sapo Challenge.
- Development of Key phrase extraction tool to generate Tag Cloud of Broadcast News stories and deployment in a real-time Multimedia Monitoring System at Voice Interaction.
- Integration of a Capitalization and Punctuation module in the end of AUDIMUS (Speech Recognizer) pipeline.
- Creation of the first BP2EP- Brazilian Portuguese to European Portuguese rule base machine translation system

Carnegie Mellon University, August 2009 – August 2010

Graduate Research Assistant

Advisors: Maxine Eskenazi and Nuno Mamede

Project: REAP.PT, REAP, and REAP Catholic Charities:

- Extending and enhancing initial version of REAP with TTS and multimedia content.
- Creation and deployment REAP Catholic Charities a system focus on teaching job vocabulary for foreign speakers in Downtown Pittsburgh.
- Leveraging text simplification tools by improving sentence splitting in difficult texts.
- Initial experiments using Amazon Mechanical Turk crowdsourcing for CALL.

Programming Languages: **Java**, **PHP**, **Javascript**, Shell Scripting, C++, Perl

Spoken Language System Laboratory of INESC-ID Lisbon,
July 2008 – July 2009

Graduate Research Assistant

Advisor: Nuno Mamede and Isabel Trancoso

Project: REAP.PT (National Portuguese Project):

- Porting and extending REAP, a Computer Assisted Language Learning System for American English, to teach European Portuguese for Foreign speakers.
- Development of Readability Metrics for European Portuguese.
- Processing and filtering large corpora using a cluster (Hadoop architecture)
- Inclusion of Speech tools (TTS) and indexed multimedia content (Broadcast News) for teaching.
- Inclusion and adaptation of Broadcast News Topic classifier
- Creation of a framework that enabled another 2 MSc thesis, a CMU-Portugal project and several research papers.
- Covered on **SIC Notícias** (portuguese cable TV), program Falar Global REAP.PT and PT-STAR. Interviews with professors Isabel Trancoso, Nuno Mamede, Luísa Coheur, and João Paulo Neto. (July, 2009)
- Summary: Porting REAP to European Portuguese
- Extensive description: MSc. Thesis – REAP.PT

Programming Languages: **Java**, **PHP**, Javascript, Shell Scripting, Python, TeX.

Spoken Language System Laboratory of INESC-ID Lisbon,
Jan 2007 – Jan 2008

Trainee Research Assistant (Undergrad)

Advisor: David Matos

Project: NLE-GRID: Natural Language Engineering on a Computational Grid (National Portuguese Project):

- Developing a rich interface to support Distributed Natural Language Tasks in a cluster of machines

Programming Languages: PHP, Javascript, XHTML.

Publications

1. Luís Marujo, Miguel Bugalho, João P. Neto, Anatole Gershman, Jaime Carbonell, **Hourly Traffic Prediction of News Stories**, 3rd International Workshop on Context-Aware Recommender Systems held as part of the 5th ACM Recommendation Systems Conference, October 2011

2. Luís Marujo, Márcio Viveiros, João P. Neto, **Keyphrase Extraction of Broadcast News**, 12th Annual Conference of the International Speech Communication Association, August 2011
3. Luis Marujo, Nuno Grazina, Tiago Luís, Wang Ling, Luísa Coheur, Isabel Trancoso, **BP2EP – Adaptation of Brazilian Portuguese Text to European Portuguese**, In Proceedings of the 15th Conference of the European Association for Machine Translation, European Association for Machine Translation, pages 129-136, Leuven, Belgium, May 2011
4. Luis Marujo, **Voting Combination of Sentences Splitting Classifiers Applied to Several Types of Texts**, Tech. Rep. 45 / 2010 INESC-ID Lisboa, September 2010
5. Luís Marujo, **REAP.PT**, Master Thesis, IST 2009
6. Luis Marujo, José Lopes, Nuno J. Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, Céu Viana, **Porting REAP to European Portuguese**, In ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009), Wroxall Abbey Estate, Warwickshire, England, September 2009
7. Luis Marujo, Wang Lin, David Martins de Matos, **Natural Language Engineering on a Computational Grid (NLE-GRID) T3 - Multi-Component Application Builder**, Tech. Rep. 33 / 2008 INESC-ID Lisboa, January 2008

**Teaching
Experience**

Carnegie Mellon University

Teaching Assistance:

- Inventing the Future of Services (Graduate course), Fall 2011
Professor: Anatole Gershman

Carnegie Mellon University, Psychology Department

Graduate course:

- Educational Goals, Instruction, and Assessment, Fall 2009
Objective: Learning how to design a course.
Project: REAP Catholic Charities

Grade: A

Professional Activities	Program committee member: - LTI Student Research Symposium, 2011 Intern member of Portuguese Order of Engineers 2011
Invited Talks	ICTI Research Presentation Luncheon, 2011, at CMU Presentation Title: “Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization
Computer Skills	Programming Languages: Java, PHP, Javascript; also experience with C++, C, Python, Matlab and Lisp Operating Systems: Mac OS X, Windows 98 - 7, UNIX, Linux (Gentoo, Fedora, OpenSuse). Hardware: Assembling PCs, LAN setup, flashing firmware, overclocking Knowledge Software: Netbeans, MS Office XP-2011, Vim, Adobe Photoshop CS, Apache Hadoop, Weka, several NLP tools, Omnigraph/MS Visio, Omniplan/MS Project, Command Line, etc.
Language Skills	Portuguese (native) English (proficient) Spanish (working knowledge) French (basic knowledge) Mandarin (very limited knowledge – about 100 characters)
Honors	Best Poster Award, S3MR - 2nd Summer School in Social Media Retrieval , for poster titled “Key-Phrase Extraction on Broadcast News” CMU-Portugal Fellowship (FCT), 2009- PetaMedia and EIT ICT Lab Grant, 2011 (to attend S3MR) FCT research fellowship, Jan 2009 – Jul 2009

Undergraduate FCT research fellowship, Jan 2007 – Jan 2008

In Honor Roll 2004 (top 10 best students out of about 10.000 students) in High School Fernando Namora, Amadora, Portugal