CURRICULUM VITAE

PERSONAL DETAILS

Name: Gustavo Henrique Paetzold

Nationality: Brazilian

Birth Date: 13/09/1991

Phone: +44 7476690262

E-mail: ghpaetzold@outlook.com
g.h.paetzold@sheffield.ac.uk

Website: https://gustavopaetzold.wordpress.com

Github: https://github.com/ghpaetzold

INTERESTS

Natural Language Processing

- Multi-Modal Multi-Lingual Text Adaptation and Simplification
- Quality Estimation for Machine Translation and Simplification

Machine Learning

- Tree Transduction
- Deep Learning
- Ensemble Methods
- Ranking Models

Compilation Methods

- Lexical Analysis
- Code Generation Optimization

LANGUAGE SKILLS

Portuguese:

- Native reading
- Native listening
- Native speaking

English:

- Fluent reading
- Fluent listening
- Fluent speaking

PROGRAMMING SKILLS

Advanced:

- Java
- Python

Intermediate:

- SQL
- Matlab
- Assembly
- C/C++
- HTML/CSS/Javascript
- NXC (Not eXactly C)
- Pure Data

CURRENT OCCUPATION

2016-Now University of Sheffield

Research Assistant in Text Adaptation

- In charge of developing novel text and workflow adaptation solutions for the Simpatico project, which aims at assisting users in accessing public administration services in various countries.
- Created a state-of-the-art approach to Lexical Simplification that employs different variations of language-agnostic context-aware word embedding models, novel unsupervised learning strategies, and new supervised Neural Network rankers.
- Developed an open-source Google Chrome extension that offers a series of free Text Adaptation solutions to the audiences being targeted in the SIMPATICO project. It allows the user to select challenging words and request complementary information about them, such as Wikipedia entries, dictionary definitions, synonyms, related words, translations and descriptive images.
- Produced various useful resources to assist in the creation of more effective Text Adaptation solutions, such as SubIMDB, a large structured corpus of subtitles for family and children that capture everyday language for non-native English speakers, the parallel Newsela dataset, composed of more than 147,000 complex-to-simple parallel sentences annotated with respect to their reading level, and the Bootstraped MRC Database, composed of automatically produced psycholinguistic features for over 85,000 words in the English vocabulary.
- Published these contributions on QATS 2016, WMT 2016, LREC 2016, NAACL 2016 and AAAI 2016.

EXPERIENCE

2010-2011 State University of Western Paraná

Assistant Lecturer in Introduction to Algorithms

- In charge of helping students during office hours.
- Granted the responsibility of grading assignments and tests.

2011-2012 State University of Western Paraná

Research Development in Robotics and Embedded Applications

- Project funded by the Brazilian Ministry of Education.
- Studied the integration between Lego robots and mobile applications.
- Research was subject for two publications in regional events.

2012-2014 State University of Western Paraná

Research Development in Natural Language and Speech Processing

- Project funded by the Brazilian Ministry of Education.

- Studied new strategies for Text Simplification and Speech Recognition.
- Research was subject for one regional publication, one national publication and several international submissions including ACL 2013.

2014 University of Sheffield

Merging of Okapi Framework and QuEst

- Project funded by the European Association for Machine Translation.
- Lead developer responsible for developing a processing step for quality estimation for the Okapi pipeline using the methods provided by QuEst.
- The project has yielded multiple international publications.

2015 University of Sheffield

Incrementing QuEst with Word-Level Features

- Project funded by the European Association for Machine Translation.
- Lead developer responsible for adding word-level features to the set of available resources of the QuEst toolkit.
- The technology produced have been employed in various Quality Estimation shared tasks and have yielded multiple international publications in conferences such as ACL.

2015 Iconic Translation Machines Ltd.

Incorporating Quality Estimation in the Iconic Translation Framework

- Project funded by the European Association for Machine Translation.
- Lead developer responsible for introducing Quality Estimation solutions into the Machine Translation framework used by the Iconic Machine Translations Ltd. company.

2016 text&form GmbH

Adapting the PET Tool for the text&form Translation Platform

- Project funded by text&form GmbH.
- Lead developer responsible for adapting the PET (Post-Editing Tool) to the specifications outlined by the text&form engineers.

EDUCATION

Ph.D. Degree in Natural Language Processing

- University of Sheffield
- Scholarship granted by the University of Sheffield
- 2014-2016

Bachelor's Degree in Computer Science

- Western Parana State University
- Score average of 93.5 points
- 2009-2013

Sandwich Graduation in Computer Science

- University of Sheffield
- Scholarship granted by the Science Without Borders program
- 2012

PUBLICATIONS

A Survey on Lexical Simplification. JAIR, Volume 60. 2017.

Complex Word Identification: Challenges in Data Annotation and System Performance. Proceedings of the 8th IJCNLP. 2017.

MASSAlign: Alignment and Annotation of Comparable Documents. Proceedings of the 8th IJCNLP. 2017.

Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs. Proceedings of the 8th IJCNLP. 2017.

A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. Proceedings of the 20th TSD. 2017.

Lexical Simplification with Neural Ranking. Proceedings of the 15th EACL. 2017.

Understanding the Lexical Simplification Needs of Non-Native Speakers of English. The 26th COLING. 2016

Collecting and Exploring Everyday Language for Predicting Psycholinguistic Properties of Words. The 26th COLING. 2016

Anita: An Intelligent Text Adaptation Tool. The 26th COLING. 2016

Multi-Level Quality Prediction with QuEst++. The 19th EAMT. 2016

SHEF-MIME: Word-level Quality Estimation Using Imitation Learning. The 1st WMT. 2016

SimpleNets: Machine Translation Quality Estimation with Resource-Light Neural Networks. The 1st WMT. 2016

SemEval 2016 Task 11: Complex Word Identification. The 10th SemEval. 2016

SV000gg at SemEval-2016 Task 11: Heavy Gauge Complex Word Identification with System Voting. The 10th SemEval. 2016

PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification. The 1st QATS. 2016

SimpleNets: Evaluating Simplifiers with Resource-Light Neural Networks. The 1st QATS. 2016

Inferring Psycholinguistic Properties of Words. The 15th NAACL. 2016

Benchmarking Lexical Simplification Systems. The 10th LREC. 2016

Unsupervised Lexical Simplification for Non-Native Speakers. The 30th AAAI. 2016

SHEF-NN: Translation Quality Estimation with Neural Networks. The 10th WMT. 2015

LEXenstein: A Framework for Lexical Simplification. The 53rd ACL. 2015

Multi-level Translation Quality Prediction with QuEst++. The 53rd ACL. 2015

Reliable Lexical Simplification for Non-Native Speakers. The 14th NAACL. 2015

EXTLex: An Extensible Lexical Analyser Capable of Detecting Lexical Errors. Revista Eletrônica Científica Inovação e Tecnologia. 2015

Okapi+QuEst: Translation Quality Estimation within Okapi. The 18th EAMT. 2015

Using Positional Suffix Trees to Perform Efficient Tree Kernel Calculation. The 20th NODALIDA. 2015

Text Simplification as Tree Transduction. The 9th STIL. 2013

EXTLex: A Configurable Lexical Analyzer. V Meeting for Computing in Parana. 2013

A Matlab® Remote Control with Voice Command Support for Lego®. V Meeting for Computing in Parana. 2013

Alternative Remote Controls for Lego® Mindstorms® NXT 2.0. IV Meeting for Computing in Parana. 2011

A Remote Control with Bluetooth® Connection for Lego®. IV Meeting for Computing in Parana. 2011

EVENTS ORGANIZED

The Complex Word Identification task of SemEval 2016

- A shared task in which participants were challenged with creating new approaches to identifying complex words for non-native English speakers.
- 21 teams participated and 42 strategies were submitted.
- http://alt.qcri.org/semeval2016/task11

SOFTWARE RELEASES

MASSAlign: Alignment and Annotation of Comparable Documents

- A framework that allows for the alignment of comparable documents at sentence and paragraph level, as well as the annotation of parallel sentences at word level.
- Developed entirely in Python.
- https://ghpaetzold.github.io/massalign

LEXenstein: A Framework for Lexical Simplification

- A complete framework that allows for the creation and evaluation of thousands of distinct Lexical Simplification strategies.
- Developed entirely in Python.
- http://ghpaetzold.github.io/LEXenstein

QuEst++: Multi-Level Quality Estimation

- Allows for one to calculate features and train Machine Learning models for Quality Estimation at word, sentence and document level.
- Developed in Java and Python.
- https://github.com/ghpaetzold/questplusplus

EXTLex: A Configurable Lexical Analyzer

- A lightweight and easy-to-use lexical analyzer that provides with a native error identification framework for the creation compilers.
- Developed entirely in Java.
- http://ghpaetzold.github.io/extlex

Morph Adorner Toolkit: Easy Text Adorning

- Provides an easy-to-use interface to the many utilities of Morph Adorner, such as noun inflection, verb tensing and syllable splitting.
- Developed entirely in Python.
- http://ghpaetzold.github.io/MorphAdornerToolkit

AWARDS

Best Computer Science Graduate of 2013

- Granted by the State University of Western Paraná.
- This award is granted annualy by the State University of Western Paraná to the student who has achieved the highest grade point average between all graduates.

Featured Student of 2013

- Granted by the Brazilian Computing Society.
- The award for Featured Student was created in 2003 by the Brazilian Computing Society to honor the best students of Computer Science in the country. Are awarded the students who are notorious for their distinct academic performance and exceptional involvement in research and extension projects.

GRANTS

LxMLS 2014 Grant

- Granted by Google.
- This grant is awarded by the sponsors to some of the participants of the yearly LxMLS event in 2014.

NAACL 2015 Student Volunteer Grant

- Granted by the NAACL 2015 sponsors.
- This grant is awarded to students who have had one or more papers accepted in the Student Research Workshop of NAACL 2015.

AAAI 2016 Student Volunteer Grant

- Granted by the AAAI 2016 sponsors.
- This grant is awarded to a few selected students who have had one or more papers accepted in AAAI 2016.

SCHOLARSHIPS

Tutorial Education Program (PET) Scholarship

- Granted by the Brazilian Ministry of Education.
- 2011-2013

Science Without Borders Scholarship

- Granted by the Brazilian Ministry of Education.
- 2012

Master's Degree Scholarship

- Granted by the Brazilian Ministry of Education.
- 2014

Doctorate Program Faculty Scholarship

- Granted by the University of Sheffield.
- 2014