WILKER FERREIRA AZIZ

July 6, 2016

Birth: 13/11/1985– São Paulo, SP – Brazil
Citizenships: Brazilian

Contact: Institute for Logic, Language and Computation
University of Amsterdam
Science Park 107, F2.11
Amsterdam, 1098 XG
Netherlands
Tel: +31 (0) 631991009
w.aziz@uva.nl
http://wilkeraziz.github.io

EDUCATION

2010–2014 **Ph.D. Computational Linguistics**
Research Institute in Information and Language Processing University of Wolverhampton
Thesis: Exact Sampling and Optimisation in Statistical Machine Translation
Supervisors: Dr. Lucia Specia (University of Sheffield)
Dr. Marc Dymetman (Xerox Research Centre in Europe)
Prof. Dr. Ruslan Mitkov (University of Wolverhampton)
Summary: In statistical machine translation, inference is performed over a high-complexity discrete distribution defined by the intersection between a translation hypergraph and a target language model. This distribution is too complex to be represented exactly and one typically resorts to approximation techniques either to perform optimisation – the task of searching for the optimum translation derivation – or sampling – the task of finding a subset of translation derivations that is statistically representative of the goal distribution. This thesis introduces an approach to exact optimisation and sampling based on a form of adaptive rejection sampling. In this view, the intractable goal distribution is upperbounded by a simpler, thus tractable, proxy distribution which is then incrementally refined to be closer to the goal until the maximum is found, or until the sampling performance exceeds a certain level.

2005–2010 **B.Sc. Computer Engineering (overall mark 83/100)**
Escola de Engenharia de São Carlos - Universidade Estadual de São Paulo (USP)
Monograph: Lexical Substitution for Statistical Machine Translation
Summary: I proposed a context model based on word co-occurrence to perform cross-language lexical substitution. I used passive-aggressive supervised learning to fit a linear model to rank translation alternatives in context. The results of this work were reported in the Cross-Language Lexical Substitution Task at SemEval-2010.
Research experience: I spent one year (from March 2009 to February 2010) at the Xerox Research Centre in Europe (Grenoble, France) where I worked on the use of context models and textual entailment to handle out-of-vocabulary words in SMT. My project was supervised by Dr. Marc Dymetman and funded by the Pascal-2 European Network of Excellence.

EMPLOYMENT

01/2015–present **Research Associate, Institute for Logic, Language and Computation, Universiteit van Amsterdam, Netherlands**
Summary: I am joining the Statistical Language Processing and Learning Lab led by Professor Khalil Sima'an in January 2015.

11/2013–12/2014 **Research Associate, Department of Computer Science, University of Sheffield, UK**
Summary: My work was funded by EPSRC under the MODIST (MOdelling DIscourse in Statistical Translation) project led by Dr. Lucia Specia. Discourse information typically requires nonlocal forms of parameterisation that go beyond the narrow context window managed by traditional decoders (such as required by a finite-state language model). I developed better decoding algorithms for SMT aiming at incorporating wider dependencies, particularly, I worked on a lazy incorporation of nonlocal parameterisation using a form of adaptive rejection sampling.

08/2013–12/2013 **Internship, Xerox Research Centre Europe (XRCE), Grenoble, France**
Summary: I worked with the Machine Learning for Document Access and Translation group under supervision of Dr. Marc Dymetman and Dr. Sriram Venkatapathy on developing an exact decoder/sampler for phrase-based SMT. Exact inference is achieved with the OS* algorithm (a technique previously developed at Xerox), a form of adaptive rejection sampling that can also be used for optimisation.

03/2010–06/2010 **Tutoring, Instituto de Ciências Matemáticas e de Computação - USP**
Summary: I gave weekly tutoring sessions on Automata Theory, Formal Language and Theory of Computation to computer science undergraduates.

03/2009–02/2010 **Internship, Xerox Research Centre Europe (XRCE), Grenoble, France**
Summary: I worked with the Cross-Language Technologies group under supervision of Dr. Marc Dymetman and Dr. Lucia Specia on the use of context models and textual entailment to improve statistical machine translation coverage and quality. My project on "Context Models for Textual Entailment and their Application to Statistical Machine Translation" was part of a project funded by Pascal-2 European Network of Excellence and it was granted the first prize of the September 2009's XRCE Intern's Day.

07/2006–12/2006 **Tutoring, Instituto de Ciências Matemáticas e de Computação - USP**
Summary: I gave weekly tutoring sessions on Linear Algebra and Ordinary Differential Equations for computer science and chemistry undergraduates.

RESEARCH INTERESTS

Automata theory and formal languages, machine learning (particularly for structured prediction). NLP applications such as machine translation, parsing and paraphrasing.

RESEARCH AND TECHNICAL SKILLS

Some experience in teaching statistical machine translation to postgraduate students.
Experience in research and development of SMT algorithms, e.g. grammar extraction (via pattern matching) and decoding (optimisation and sampling), MT evaluation as well as other NLP applications such as sentence- and word-alignment, word-sense disambiguation, lexical substitution and paraphrasing through pivoting.
Experience in exploiting machine learning techniques to model natural language tasks such as named-entity recognition and semantic role labelling.

Programming languages: C/C++, Python, Perl, Java.

Code: github (wilkeraziz), bitbucket (wilkeraziz)

Language skills: Portuguese (native), English (fluent), French (basic) and Italian (basic).

AWARDS AND SCHOLARSHIPS

RIILP (UK) PhD scholarship (10/2010–10/2013)

XRCE best internship award (09/2009)

FAPESP (Brazil) scientific initiation scholarship (02/2008–02/2009)

FAPESP (Brazil) scientific initiation scholarship (01/2007–01/2008)

INVITED TALKS

04/2014 Invited talk, University of Amsterdam, Amsterdam, The Netherlands
Title: Exact Inference for Statistical Machine Translation
Summary: In this presentation I talk about exact decoding and unbiased sampling for hierarchical and phrase-based SMT based on a coarse-to-fine strategy. In this view the intractable intersection between the translation forest and the language model is replaced by a simpler, thus tractable, intersection with a lower-order upperbound on the true LM distribution. The resulting distribution is then incrementally refined in an adaptive rejection sampling fashion.

05/2013 Invited talk, University of Sheffield, Sheffield, UK
Title: Exact Optimisation and Sampling for Statistical Machine Translation
Summary: Preliminary findings of my PhD where I present the OS* algorithm (Dymetman et al, 2012) and how this algorithm can be used to perform exact optimisation and sampling for SMT.

09/2011 Tutorial, RANLP, Hissar, Bulgaria
Summary: Together with Lucia Specia I gave a 3-hour tutorial on SMT.

10/2011 Invited Talk, Universidade de São Paulo, São Carlos, Brazil
Title: Improving Chunk-based Semantic Role Labelling with Lexical Features
Summary: Findings of my investigation on improving semantic role labelling by using lexical features published at RANLP-2011.

01/2010 Invited talk, University of Wolverhampton, Wolverhampton, UK
Title: Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-Vocabulary Words
Summary: Findings of the work I developed at XRCE on handling unknown words using Textual Entailment and incorporating an expert model into a standard SMT system.

PUBLICATIONS

[1] Wilker Aziz. Grasp: Randomised semiring parsing. *The Prague Bulletin of Mathematical Linguistics*, 104(1):51–62, October 2015.

[2] Wilker Aziz, Marc Dymetman, and Lucia Specia. Exact decoding for phrase-based statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1237–1249, Doha, Qatar, October 2014. Association for Computational Linguistics.

[3] Wilker Ferreira Aziz. *Exact Sampling and Optimisation in Statistical Machine Translation*. PhD thesis, University of Wolverhampton, 2014.

[4] Wilker Aziz, Maarit Koponen, and Lucia Specia. Sub-sentence level analysis of machine translation post-editing effort. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors, *Post-editing of Machine Translation: Processes and Applications*, chapter 8. Cambridge Scholars Publishing, 2014.

[5] Wilker Aziz, Marc Dymetman, and Sriram Venkatapathy. Investigations in exact inference for hierarchical translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 472–483, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Wilker Aziz and Lucia Specia. Multilingual wsd-like constraints for paraphrase extraction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 202–211, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] Wilker Aziz, Ruslan Mitkov, and Lucia Specia. Ranking machine translation systems via post-editing. In *In Proceedings of Text, Speech and Dialogue (TSD)*, volume 8082 of *Lecture Notes in Computer Science*, pages 410–418, Pilsen, Czech Republic, September 2013. Springer Berlin Heidelberg.

[8] Luciana Ramos Maarit Koponen, Wilker Aziz and Lucia Specia. Post-editing time as a measure of cognitive effort. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 11–20, San Diego, USA, October 2012. Association for Machine Translation in the Americas (AMTA).

[9] Miguel Rios, Wilker Aziz, and Lucia Specia. UOW: Semantically informed text similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 673–678, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[10] Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[11] Wilker Aziz and Lucia Specia. PET: a tool for post-editing and assessing machine translation. In *The 16th Annual Conference of the European Association for Machine Translation*, EAMT '12, page 99, Trento, Italy, May 2012.

[12] Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. Cross-lingual sentence compression for subtitles. In *The 16th Annual Conference of the European Association for Machine Translation*, EAMT '12, pages 103–110, Trento, Italy, May 2012.

[13] Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. Predicting machine translation adequacy. In *Proceedings of the 13th Machine Translation Summit*, pages 513–520, Xiamen, China, September 2011.

[14] Wilker Aziz, Miguel Rios, and Lucia Specia. Shallow semantic trees for smt. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 316–322, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[15] Miguel Rios, Wilker Aziz, and Lucia Specia. TINE: A metric to assess mt adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

[16] Wilker Aziz, Miguel Rios, and Lucia Specia. Improving chunk-based semantic role labeling with lexical features. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 226–232, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee.

[17] Sheila C. M. de Sousa, Wilker Aziz, and Lucia Specia. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee.

[18] Wilker Aziz and Lucia Specia. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, Cuiabá, MT, Obtober 2011.

[19] Wilker Aziz, Marc Dymetman, Shachar Mirkin, Lucia Specia, Nicola Cancedda, and Ido Dagan. Learning an expert from human annotations in statistical machine translation: the case of out-of-vocabulary words. In *14th Annual Conference of the European Association for Machine Translation*, EAMT '10, pages 28–35, Saint-Raphael, France, 2010.

[20] Wilker Aziz and Lucia Specia. USPwlv and WLVusp: Combining dictionaries and contextual information for cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 117–122, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

PATENTS

U.S. Patent Application Filing: SAMPLING AND OPTIMIZATION IN PHRASED-BASED MACHINE TRANSLATION USING AN ENRICHED LANGUAGE MODEL REPRESENTATION
Inventor(s): Marc Dymetman; Wilker Aziz; Sriram Venkatapathy
U.S. Ser. No.: 13/750,338
Filed on: 01/25/2013

U.S. Patent Application Filing: DYNAMIC BI-PHRASES FOR STATISTICAL MACHINE TRANSLATION
Inventor(s): Marc Dymetman; Wilker Aziz; Nicola Cancedda; Jean-Marc Coursimault; Vassilina Nikoulina; Lucia Specia.
U.S. Ser. No.: 12/780,040
Filed on: 05/20/2010