WILKER FERREIRA AZIZ

May 14, 2017

Birth: 13/11/1985– São Paulo, SP – Brazil
Languages: Portuguese (native), English (fluent), Italian and French (basic)

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
https://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: Prof. Dr. Lucia Specia (University of Sheffield)
Dr. Marc Dymetman (Xerox Research Centre in Europe)
Prof. Dr. Ruslan Mitkov (University of Wolverhampton)
Summary: I introduce an approach to exact optimisation and sampling based on a form of adaptive rejection sampling which addresses challenges in global optimisation and unbiased sampling in high-dimensional discrete spaces. In this view, an intractable goal distribution is upperbounded by a tractable proxy distribution which is then incrementally refined to be closer to the goal.

2005–2010 **B.Sc. Computer Engineering**
Escola de Engenharia de São Carlos - Universidade Estadual de São Paulo (USP)
Monograph: Lexical Substitution for Statistical Machine Translation
Summary: I propose a context model based on word co-occurrence and supervised learning to rank for cross-language lexical substitution.

EMPLOYMENT

01/2015–present **Research Associate, Institute for Logic, Language and Computation, Universiteit van Amsterdam, Netherlands**
Summary: I joined the Statistical Language Processing and Learning Lab led by Professor Khalil Sima'an in January 2015 where I work on several aspects of machine translation (e.g. word alignment, word reordering, and morphological analysis and generation) and paraphrasing employing log-linear, Bayesian, and neural network models.

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. I developed better decoding algorithms for SMT aiming at incorporating global features, 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.

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.

RESEARCH INTERESTS

My work focusses on natural language processing applications (e.g. machine translation, parsing and paraphrasing) with a special focus on unsupervised learning of language structure. My research interests and experience span several disciplines including automata theory and formal grammars, machine learning, Bayesian statistics, deep generative models, and global optimisation.

TEACHING

I have extensive teaching experience on various topics:

- probabilistic graphical models
- Bayesian methods for NLP
- approximate probabilistic inference: Markov chain Monte Carlo sampling and variational inference
- weighted automata and grammars, semirings, and deductive systems
- statistical and neural approaches to machine translation

Since 2015, I coordinate, design, and implement MSc courses offered at UvA.

Course	**Programme**	**Offered**	**Role**
Natural Language Processing 2	Master of AI	Spring (2015, 2016, 2017)	Coordinator
Description		`https://uva-slpl.github.io/nlp2/`	
Statistical and neural approaches to natural language processing applications such as word alignment, synchronous parsing, machine translation, and paraphrasing. The course focusses on unsupervised learning with probabilistic models covering frequentist, Bayesian, and neural methods.			
Bayesian inference for PCFGs	Master of Logic	Winter 2017	Coordinator
Description		`https://wilkeraziz.github.io/teaching.html`	
Unsupervised learning of probabilistic context-free grammars with Dirichlet priors and Bayesian inference.			
Monte Carlo sampling for PCFGs	Master of Logic	Summer 2015	Coordinator
Description		`https://wilkeraziz.github.io/teaching.html`	
Parsing as weighted deduction and sampling algorithms for probabilistic context-free grammars.			

For a complete list of courses and projects including available material and project outcomes, please refer to `https://wilkeraziz.github.io/teaching.html`.

SUPERVISION

I have extensive experience supervising students at all levels.

Name	**Programme**	**Status**	**Role**	**Thesis title**
Joost Bastings	PhD	2nd year	co-supervisor	Incorporating linguistic structure into neural machine translation
Joachim Daiber	PhD	3rd year	co-supervisor	Linguistic typology and machine translation: understanding and exploiting differences and similarities between languages
Philip Schulz	PhD	3rd year	co-supervisor	Hierarchical Bayesian models for machine translation and word alignment
Miloš Stanojević	PhD	final year	co-supervisor	Permutation forests for modelling word order in machine translation
Sander Bijl de Vroe	Master of AI	final year	co-supervisor	Deep generative models of morphological segmentation and analysis for neural machine translation
Guido Linders	Bachelor of AI	complete (2016)	supervisor	Feature-rich unsupervised word alignment models
Iason de Bondt	Bachelor of AI	complete (2015)	supervisor	Sampling from probabilistic context-free grammars

For more information and links to theses, please refer to `https://wilkeraziz.github.io/people.html`.

PUBLICATIONS

For e-prints, slides, talks, data, and code, refer to `https://wilkeraziz.github.io/publications.html`. For invited talks, lectures, and tutorials, refer to `https://wilkeraziz.github.io/talks.html`.

Selected publications representative of my current interests

[1] *Graph convolutional encoders for syntax-aware neural machine translation.*
`https://arxiv.org/pdf/1704.04675.pdf`

[2] *Fast collocation-based Bayesian HMM word alignment*
`http://www.aclweb.org/anthology/C/C16/C16-1296.pdf`

Complete list

[1] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*, 2017. Under review at EMNLP17.

[2] Philip Schulz and Wilker Aziz. Fast collocation-based Bayesian HMM word alignment. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3146–3155, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[3] Joachim Daiber, Miloš Stanojević, Wilker Aziz, and Khalil Sima'an. Examining the relationship between preordering and word order freedom in machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 118–130, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Philip Schulz, Wilker Aziz, and Khalil Sima'an. Word alignment without null words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 169–174, Berlin, Germany, August 2016. Association for Computational Linguistics.

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

[6] Raymond W. M. Ng, Kashif Shah, Wilker Aziz, Lucia Specia, and Thomas Hain. Quality estimation for asr k-best list rescoring in spoken language translation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015.*, pages 5226–5230, South Brisbane, Queensland, Australia, April 2015.

[7] 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.

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

[9] Raymond W. M. Ng, Mortaza Doulaty, Rama Doddipatla, Wilker Aziz, Kashif Shah, Oscar Saz, Madina Hasan, Ghada AlHarbi, Lucia Specia, and Thomas Hain. The USFD spoken language translation system for IWSLT 2014. *CoRR*, abs/1509.03870, 2015.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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).

[15] 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.

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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