Kenneth Heafield

to@kheafield.com https://kheafield.com

INTERESTS Neural machine translation, big data, language modeling, and natural language processing

EDUCATION

PhD, Carnegie Mellon 8/2008–9/2013
Efficient Language Modeling Algorithms with Applications to Statistical Machine Translation
Department: Language Technologies Institute in the School of Computer Science
Adviser: Professor Alon Lavie

Bachelor of Science, Caltech 9/2003–3/2007
Double major in Mathematics and Computer Science, with honors.

EXPERIENCE

Lecturer (Assistant Professor), University of Edinburgh 8/2015–Present
Lead research in neural machine translation, language modeling, and algorithms. Coordinating European Union projects, supervising 3 PhD students, and 3 postdocs.

Senior Research Scientist, Bloomberg 8/2014–7/2015
Machine translation lead. Created new models and investigated hardware acceleration for publication. Solicitied, reviewed, and advocated funding for research proposals from universities.

Postdoctoral Scholar, Stanford 10/2013–7/2014
Responsible for machine translation efforts at Stanford, including supervising two PhD students and four Master's students. Research included web-scale text processing, algorithms for machine translation, and applications of neural networks.

Research Associate, University of Edinburgh 8/2011–12/2011; 8/2012–9/2013
Created an efficient search algorithm for syntactic machine translation, made language model estimation efficient, contributed to the Moses machine translation system, and informally advised PhD students.

Software Engineer, Google 3/2007–8/2008
Optimized language classification for card catalog information about books as part of the Google Books team. Created the ranking function for a search system in Picasa Web Albums. Lectured at MIT about Hadoop.

Intern, Infosys Technologies 7/2006–9/2006
Travelled to Bangalore, India for an internship with the Software Engineering Technology Lab. Applied latent Dirichlet allocation to automatically organize source code.

Undergraduate Researcher, Netlab at Caltech 6/2005–6/2006
Developed an error model for kernel principal component analysis (kPCA) and applied it to automatically analyze computer network traffic and flag possible attacks.

Undergraduate Researcher, Galaxy Evolution Explorer 6/2004–3/2007
Found variable stars by creating and mining a 193-million row database of measurements from a satellite.

GRANTS RECEIVED

EuroPat: Unleashing European Patent Translations *Coordinator* EU: €695,890 2019

Continued Web-Scale Provision of Parallel Corpora for European Languages EU: €889,649 2019

UKRI Centre for Doctoral Training in Natural Language Processing *Industry Lead*	EPSRC: £6,530,681	2019
Mining non-English Parallel Corpora	eBay: $30,000	2018
Faster Marian on the CPU	eBay: $30,000	2018
Compression in Distributed Machine Translation	EPSRC: 150,000 GPU hours	2018
Bergamot: Browser-based Multilingual Translation *Coordinator*	EU: €2,999,096	2018
Broader Web-Scale Provision of Parallel Corpora for European Languages	EU: €907,976	2018
ParaCrawl	EPCC: 500,000 CPU hours	2018
Fast Neurons on Xeons	Intel: £28,302	2018
Distributed Machine Translation	EPSRC: 150,000 GPU hours	2017
System for cross-language information processing, translation, and summarization	IARPA: $1,245,515	2017
Provision of web-scale parallel corpora for official European languages	EU: €585,414	2017
Medical machine translation	EPSRC: £51,993	2017
Scalable recurrent neural networks	US DOE: 9,000,000 supercomputer hours	2017
Making web crawl a Turing resource	Alan Turing Institute: £26,137	2017
Cloud computing for MSc dissertations	Google Cloud: $17,500	2017
Open data: mining translations and transcripts from the web	Mozilla: $64,143	2017
Training neural machine translation systems	Microsoft Azure: $750,000	2017
Mining and training on translations from the web	Microsoft Azure: $20,000	2017
Neural network primitives and distributed training	Intel: £28,302	2017
Phrase-based decoding	eBay: $30,000	2016
Decoding methods for spelling correction and synonym generation	Facebook: $50,000	2016
Local coarse-to-fine decoding for long-distance models	Google: $57,724	2015
Faster decoding and better features via local coarse-to-fine	Amazon: $70,949	2015
Faster machine translation and principled rule learning	Bloomberg: $150,000	2014
Applying tera-scale language models to advance machine translation	NSF: 200,000 CPU hours	2013

GRANTS FUNDED

Christopher Manning. Natural Language Processing and Machine Learning. $75,000 gift from Bloomberg, 2015.

Shay Cohen. Latent-Variable Learning for Transition-Based Parsing. $63,379 gift from Bloomberg, 2015.

Philipp Koehn. High Quality Parallel Corpus Extraction from the Web. $50,000 gift from Bloomberg, 2015.

Lane Schwartz. US Machine Translation Marathon at UIUC. $10,000 gift from Bloomberg, 2015.

Fei-Fei Li. SAILORS AI summer outreach. $10,000 gift from Bloomberg, 2015.

STUDENTS

PhD: Alham Fikri Aji, Maximiliana Behnke, Anna Currey
PhD as assistant supervisor: Nikolay Bogoychev, Naums Mogers, Amna Shahab
PhD examiner: Dominik Wurzer, 2017
MPhil (with Miles Osborne): Luke Shrimpton, 2016
MSc: supervised nine theses

OPEN-SOURCE SOFTWARE

KenLM
An efficient library for estimating and querying language models. Compared with SRILM, querying is 2.4 times as fast and uses 57% of the memory. It has been adopted by all major open-source machine translation systems.

Hypergraph Search
Implements my new search algorithm for syntactic machine translation, which makes translation 1.6–6.0 times as fast as with cube pruning.

System Combination (MEMT)
Combines the outputs of multiple machine translation systems into a single sentence with better quality.

CONFERENCE PAPERS

Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, **Kenneth Heafield**, Antonio Valerio Miceli Barone, and Rico Sennrich. The University of Edinburgh's Submissions to the WMT18 News Translation Task. *EMNLP 2018 Third Conference on Machine Translation (WMT18)*, Brussels, Belgium, October, 2018.

Anna Currey and **Kenneth Heafield**. Multi-Source Syntactic Neural Machine Translation. *2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November, 2018.

Nikolay Bogoychev, **Kenneth Heafield**, Alham Fikri Aji, and Marcin Junczys-Dowmunt. Accelerating Asynchronous Stochastic Gradient Descent for Neural Machine Translation. *2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November, 2018.

Marcin Junczys-Dowmunt, **Kenneth Heafield**, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. Marian: Cost-effective High-Quality Neural Machine Translation in C++. *2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, July, 2018.

Roman Grundkiewicz and **Kenneth Heafield**. Neural Machine Translation Techniques for Named Entity Transliteration. *The Seventh Named Entities Workshop (NEWS)*, Melbourne,

Australia, July, 2018.

Hieu Hoang, Tomasz Dwojak, Rihards Krislauks, Daniel Torregrosa, and **Kenneth Heafield**. Fast Neural Machine Translation Implementation. *2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, July, 2018.

Anna Currey and **Kenneth Heafield**. Unsupervised Source Hierarchies for Low-Resource Neural Machine Translation. *Relevance of Linguistic Structure in Neural NLP*, Melbourne, Australia, July, 2018.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, **Kenneth Heafield**, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast Neural Machine Translation in C++. *56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, July, 2018.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and **Kenneth Heafield**. Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task. *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, June, 2018.

Alham Fikri Aji and **Kenneth Heafield**. Sparse Communication for Distributed Gradient Descent. *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September, 2017.

Kenneth Heafield, Chase Geigle, Sean Massung, and Lane Schwartz. Normalized Log-Linear Language Model Interpolation is Efficient. *The 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, August, 2016.

Kenneth Heafield, Rohan Kshirsagar, and Santiago Barona. Language Identification and Modeling in Specialized Hardware. *The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, Beijing, China, July, 2015.

Kenneth Heafield, Michael Kayser, and Christopher D. Manning. Faster Phrase-Based Decoding by Refining Feature State. *Association for Computational Linguistics*, Baltimore, MD, USA, June, 2014.

Christian Buck, **Kenneth Heafield**, and Bas van Ooyen. N-gram Counts and Language Models from the Common Crawl. *Language Resources and Evaluation Conference*, Reykjavík, Iceland, May, 2014.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable Modified Kneser-Ney Language Model Estimation. *51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August, 2013.

Kenneth Heafield, Philipp Koehn, and Alon Lavie. Grouping Language Model Boundary Words to Speed K-Best Extraction from Hypergraphs. *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, June, 2013.

Kenneth Heafield, Philipp Koehn, and Alon Lavie. Language Model Rest Costs and Space-Efficient Storage. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, July, 2012.

Kenneth Heafield and Alon Lavie. Voting on N-grams for Machine Translation System Combination. *Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA, November, 2010.

Girish Maskeri, Santonu Sarkar, and **Kenneth Heafield**. Mining Business Topics in Source Code using Latent Dirichlet Allocation. *1st India Software Engineering Conference*, Hyderabad, India, February, 2008. ***10-year test of time award.***

Stanley Browne, Jonathan Wheatley, Barry Welsh, Mark Seibert, **Kenneth Heafield**, R. Michael Rich, and the GALEX Science Team. RR Lyrae Stars in the Far Ultraviolet: GALEX Observations Compared with Theoretical Predictions. *American Astronomical Society 207th Meeting*, Washington, DC, USA, June, 2006.

Barry Welsh, Jonathan Wheatley, **Kenneth Heafield**, Mark Seibert, Stanley Browne, and the GALEX Science Team. The Flaring UV Sky. *American Astronomical Society 205th Meeting*, San Diego, California, USA, January, 2005.

Refereed Workshop Papers

Anna Currey, Antonio Valerio Miceli Barone, and **Kenneth Heafield**. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. *EMNLP 2017 Second Conference on Machine Translation (WMT17)*, Copenhagen, Denmark, September, 2017.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, **Kenneth Heafield**, Antonio Valerio Miceli Barone, and Philip Williams. The University of Edinburgh's Neural MT Systems for WMT17. *EMNLP 2017 Second Conference on Machine Translation (WMT17)*, Copenhagen, Denmark, September, 2017.

Nadir Durrani, Barry Haddow, Philipp Koehn, and **Kenneth Heafield**. Edinburgh's Phrase-based Machine Translation Systems for WMT-14. *ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June, 2014.

Julia Neidert, Sebastian Schuster, Spence Green, **Kenneth Heafield**, and Christopher D. Manning. Stanford University's Submissions to the WMT 2014 Translation Task. *ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June, 2014.

Nadir Durrani, Barry Haddow, **Kenneth Heafield**, and Philipp Koehn. Edinburgh's Machine Translation Systems for European Language Pairs. *ACL 2013 Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August, 2013.

Kenneth Heafield, Hieu Hoang, Philipp Koehn, Tetsuo Kiso, and Marcello Federico. Left Language Model State for Syntactic Machine Translation. *International Workshop on Spoken Language Translation*, San Francisco, California, USA, December, 2011.

Kenneth Heafield. KenLM: Faster and Smaller Language Model Queries. *EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July, 2011.

Kenneth Heafield and Alon Lavie. CMU System Combination in WMT 2011. *EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July, 2011.

Kenneth Heafield and Alon Lavie. CMU Multi-Engine Machine Translation for WMT 2010. *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, July, 2010.

Kenneth Heafield, Greg Hanneman, and Alon Lavie. Machine Translation System Combination with Flexible Word Ordering. *EACL 2009 Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March, 2009.

Journal Articles

Kenneth Heafield and Alon Lavie. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. *The Prague Bulletin of Mathematical Linguistics* 93. January, 2010.

Jonathan H. Clark, Jonathan Weese, Byung Gyu Ahn, Andreas Zollmann, Qin Gao, **Kenneth Heafield**, and Alon Lavie. The Machine Translation Toolpack for LoonyBin: Automated Management of Experimental Machine Translation HyperWorkflows. *The Prague Bulletin of Mathematical Linguistics* 93. January, 2010.

Barry Welsh, Johathan Wheatley, **Kenneth Heafield**, Mark Seibert, and the GALEX Science Team. The GALEX Ultraviolet Variability Catalog. *The Astronomical Journal* 130. 2005.

Patents

Girish Maskeri Rama, **Kenneth Heafield**, and Santonu Sarkar. Identification of Topics in Source Code. US Patent 8209665 filed in 2009 and issued June, 2012.

Taylor Curtis and **Kenneth Heafield**. Systems and Methods for Identifying Similar Documents. US Patent 7958136 filed in 2008 and issued June, 2011.

Invited Talks

Intel 2018
Marian Machine Translation

Microsoft 2018
Sharpening Machine Translation Decoding

European Language Resource Consortium 2017
Translation quality, neural machine translation and language resources

Georgetown 2017
Machine Translation is Too Slow

Facebook Paris 2017
Translation and Distributed Training

Google Mountain View 2016
WMT Tricks and Normalization

Facebook Menlo Park 2016
Move Fast and Normalize in Machine Translation

Facebook London Faculty Summit 2015
Language Modeling at Web Scale

Microsoft Research 2014
Scalable High-Quality Language Modeling and Machine Translation

Facebook 2014
Scalable High-Quality Language Modeling and Machine Translation

University of Edinburgh 2014
Scalable High-Quality Language Modeling and Machine Translation

Bloomberg 2013
Faster and Better Machine Translation

Google Mountain View 2013
Language Model Algorithms

Apple 2013
Language Model Algorithms

Numen Digital 2013
Faster Decoding for Machine Translation and Lattices

Xerox Research Centre Europe 2013
Faster Decoding for Machine Translation and Lattices

Qatar Computing Research Institute and Carnegie Mellon-Qatar 2013
Faster Search for Machine Translation

Hong Kong University of Science and Technology 2012
Language Model Rest Costs and Space-Efficient Storage

Tutorials

Language Modeling, Machine Translation Marathon 2011–2017

Language Model Implementation, Machine Translation Marathon 9/2013

Language Modeling with KenLM, Qatar Computing Research Institute 3/2013

Chart Based Decoding, Machine Translation Marathon 9/2012

Teaching

Extreme Computing, University of Edinburgh Fall 2016, Fall 2017

Computer Programming Skills and Concepts, University of Edinburgh Fall 2016

Extreme Computing, University of Edinburgh Fall 2015

Guest Course Lecture: Machine Translation, Carnegie Mellon 3/2013

Guest Course Lecture: Advanced NLP, University of Edinburgh 10/2012

Teaching Assistant: Language and Statistics, Carnegie Mellon Spring 2012

Teaching Assistant: Algorithms for NLP, Carnegie Mellon Fall 2010

Lecturer: Introduction to Hadoop, MIT 1/2008

Awards

Bloomberg BFIRST 2014
Bloomberg award for Knowledge Discovery and Data Mining (KDD)

Student Travel Grant 2012
$800 in travel funded by the EMNLP conference

National Science Foundation Graduate Research Fellowship 2008–11
$121,500 in stipend and tuition over three years

Google Peer Bonus and Site Award 2008
For lecturing at MIT on Hadoop while a Software Engineer at Google

International Collegiate Programming Contest Regional 2006–2007
Ranked third of fifty in a team of two instead of three

Carnation Scholarship 2005–06
Year of full Caltech tuition based on academic merit; 38 awarded per year

Richard and Dena Krown Summer Undergraduate Research Fellowship 2005
$5,000 for ten weeks of summer research in networking

Summer Undergraduate Research Fellowship 2004
$5,000 for ten weeks of summer research in astronomy data mininug

Program Committees	Transactions of the ACL (TACL) reviewing team	May 2016–Present
	Association for Computational Linguistics (ACL)	2014–2016,2018
	Empirical Methods in Natural Language Processing (EMNLP)	2012–2018
	Workshop on Statistical Machine Translation (WMT)	2011–2017
	North American Association for Computational Linguistics (NAACL)	2013–2014, 2016, 2018
	European Association for Computational Linguistics (EACL)	2012, 2017
	International Joint Conference on Artificial Intelligence (IJCAI)	2017
	International Conference on Computational Linguistics (COLING)	2012, 2014, 2018
	Transactions on Asian Language Information Processing (TALIP)	2011, 2014, 2015
	Machine Translation Journal	2011