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. Supervising four PhD students.

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

Kenneth Heafield. Training neural machine translation systems. $20,000 in Azure credits, 2017.

Kenneth Heafield. Neural network primitives and distributed training. 40% buyout, 2017.

Kenneth Heafield. Phrase-based decoding. $30,000 gift from eBay, 2016.

Kenneth Heafield. Decoding methods for spelling correction and synonym generation. $50,000 gift from Facebook's Research and Academic Relations Program, 2016.

Kenneth Heafield. Local coarse-to-fine decoding for long-distance Models. $57,724 gift from Google Research Awards, 2015.

Kenneth Heafield and Rico Sennrich. Faster decoding and better features via local coarse-to-fine. $70,949 gift from Amazon Academic Research Awards, 2015.

Kenneth Heafield, and Christopher Manning. Faster machine translation and principled rule learning. $150,000 gift from Bloomberg, 2014.

Christopher D. Manning, **Kenneth Heafield**, and Dan Jurafsky. Applying tera-scale language models to advance machine translation. National Science Foundation XSEDE grant CCR140009. These grants award time on supercomputers, in this case 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.

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.

Community Evaluation Results

First Place in Three Language Pairs, Workshop on Machine Translation 2013
Using a language model estimated on 126 billion tokens, my three submissions ranked first in their respective language pairs, each of which had 11–13 participants.

First Place System Combinations, Workshop on Machine Translation 2011
Submitted to all ten system combination tracks, each of which had 2–8 participants. Eight of my submissions ranked first in their respective tracks.

Conference Papers

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.

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

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

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 Language Modeling at Web Scale	2015
	Microsoft Research Scalable High-Quality Language Modeling and Machine Translation	2014
	Facebook Scalable High-Quality Language Modeling and Machine Translation	2014
	University of Edinburgh Scalable High-Quality Language Modeling and Machine Translation	2014
	Bloomberg Faster and Better Machine Translation	2013
	Google Mountain View Language Model Algorithms	2013
	Apple Language Model Algorithms	2013
	Numen Digital Faster Decoding for Machine Translation and Lattices	2013
	Xerox Research Centre Europe Faster Decoding for Machine Translation and Lattices	2013
	Qatar Computing Research Institute and Carnegie Mellon-Qatar Faster Search for Machine Translation	2013
	Hong Kong University of Science and Technology Language Model Rest Costs and Space-Efficient Storage	2012
Tutorials	Language Modeling, Machine Translation Marathon	9/2014
	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
	Language Modeling, Machine Translation Marathon	9/2011
Teaching	Extreme Computing, 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** Bloomberg award for Knowledge Discovery and Data Mining (KDD)	2014
	Student Travel Grant $800 in travel funded by the EMNLP conference	2012
	National Science Foundation Graduate Research Fellowship $121,500 in stipend and tuition over three years	2008–11

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–June 2018
Association for Computational Linguistics (ACL)	2014–2016
Empirical Methods in Natural Language Processing (EMNLP)	2012–2016
Workshop on Statistical Machine Translation (WMT)	2011–2016
North American Association for Computational Linguistics (NAACL)	2013–2014, 2016
International Conference on Computational Linguistics (COLING)	2012, 2014
European Association for Computational Linguistics (EACL)	2012
Transactions on Asian Language Information Processing (TALIP)	2011, 2014, 2015
Machine Translation Journal	2011