

Kenneth Heafield

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave GHC 5407
Pittsburgh, PA 15213

<cv@kheafield.com>
<http://kheafield.com>

Interests

Machine translation, machine learning, distributed systems, theoretical computer science

Education

PhD program, Carnegie Mellon

August 2008–

Language Technologies Institute in the School of Computer Science; 3.9/4.0 GPA.

Advised by Alon Lavie, I work on efficient machine translation and system combination.

- Wrote KenLM, an efficient open-source language model library. Compared with the widely-used SRILM, KenLM's default is 2.4 times as fast while using 57% of the memory. Additional options save more memory. It is used by several translation systems: Moses, cdec, Joshua, and Ncode.
- Won the Workshop on Machine Translation (WMT) 2011 system combination task in eight of ten language pairs. In WMT 2010, won six of eight language pairs. My code, dubbed MEMT (Multi-Engine Machine Translation), is open-source.
- Visited Philipp Koehn at the University of Edinburgh August–December 2011 to improve language modeling in the Moses translation system.

Bachelor of Science, Caltech

September 2003–March 2007

Double major in Mathematics and Computer Science; 3.8/4.0 GPA, with honors.

- Courses focused on formal language theory, distributed systems, information theory, and combinatorics.
- Went to Bangalore for a summer internship with Infosys.
- Worked for two Caltech research groups: Netlab and Galaxy Evolution Explorer.
- Finished a quarter early and went to work for Google.

Skills

Languages Extensive C++, C, Ruby, SQL, Bash, and L^AT_EX; Some Java, HTML, and CSS

Software Contributed to Moses, cdec, and Joshua; Taught Hadoop; Extensive Boost and STL; Administered Linux, PostgreSQL, and Apache; Used MySQL, Octave, Gnuplot, and PBS

Awards

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

2006–07

Ranked third of fifty in a team of two instead of three

Carnation Scholarship

2005–06

Full Caltech tuition academic merit scholarship, 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

Employment Experience

Google <http://books.google.com> March 2007–August 2008

As a Software Engineer with Google Book Search, I worked on a team that uses machine learning to compile card catalogs from multiple sources into a single coherent catalog of books. Previously, I created the scoring system behind a search function in Picasa Web Albums. To share Google’s approach to distributed systems, I lectured at MIT on the Hadoop MapReduce framework.

Infosys Technologies <http://www.infosys.com> July–September 2006

I traveled to Bangalore, India to intern with the research division of Infosys, India’s second largest software outsourcing company. We investigated automatic reorganization of legacy source code. Specifically, I applied and customized Latent Dirichlet Allocation to derive topics from names of functions and local variables. For example, it found SSL and logging topics in Apache source code while correctly tagging files belonging to both topics.

Netlab <http://netlab.caltech.edu> June 2005–June 2006

As a Richard and Dena Krown Summer Undergraduate Research Fellow, I developed an error model for kernel Principal Component Analysis (kPCA). Professor Low hired me to continue with implementation during the school year. I applied it to identify possible attacks in network traffic, which appear as points with unusually high distance from the manifold learned by kPCA.

Fastsoft <http://www.fastsoft.com> January–April 2006

Netlab spun off a startup and I worked for them as a contractor. Using FAST TCP, the Netlab algorithm responsible for breaking Internet speed records, their Aria product accelerates connections passing through it. This allows senders to use high performance networks more efficiently without custom operating systems. I setup experiments and worked on the performance monitoring and configuration interface.

Galaxy Evolution Explorer <http://www.galex.caltech.edu> June 2004–March 2007

I started working for the Galaxy Evolution Explorer (GALEX) project as a Summer Undergraduate Research Fellow. My goal was finding variable stars and asteroids in observations made by their satellite. To do so, I created a database of all 193 million source measurements and used it to find and analyze over ninety variable objects. The findings were reported in two posters and one journal article. After the summer, they hired me to continue working on the database and to help scientists find interesting data.

Publications

Paper and Poster Heafield, Hoang, Koehn, Kiso, and Federico. Left Language Model State for Syntactic Machine Translation. Proc. International Workshop on Spoken Language Translation, San Francisco, CA, December 8–9, 2011.

Paper and Presentation Heafield, 2011. KenLM: Faster and Smaller Language Model Queries. Proc. EMNLP 2011 Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, July 30–31, 2011.

Paper and Poster Heafield and Lavie, 2011. CMU System Combination in WMT 2011. Proc. EMNLP 2011 Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, July 30–31, 2011.

Paper and Poster Heafield and Lavie, 2010. Voting on N-grams for Machine Translation System Combination. Proc. Ninth Conference of the Association for Machine Translation in the Americas, Denver, Colorado, October 31–November 5.

Paper and Poster Heafield and Lavie, 2010. CMU Multi-Engine Machine Translation for

WMT 2010. Proc. ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, Uppsala, Sweden, July 15–16.

Paper and Presentation Heafield and Lavie, 2010. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. The Prague Bulletin of Mathematical Linguistics 93, pages 27–36. ISBN 978-80-904175-4-0. doi: 10.2478/v10108-010-0008-4.

Presentation Heafield, 2009. CMU-StatXfer Group System Combination. Proc. NIST Open MT Workshop 2009 at MT Summit XII, Ottawa, Canada, August 31–September 1.

Paper and Poster Heafield, Hanneman, and Lavie, 2009. Machine Translation System Combination with Flexible Word Ordering. Proc. EACL 2009 Fourth Workshop on Statistical Machine Translation, Athens, Greece, March 30–31.

Patent Application Rama, Heafield, and Sarkar, 2009. Identification of Topics in Source Code. US patent application number 20090254884. Indian patent application 877/CHE/2008.

Paper and Presentation Rama, Sarkar, and Heafield, 2008. Mining Business Topics in Source Code using Latent Dirichlet Allocation. Proc. 1st India Software Engineering Conference, pages 113–120, Hyderabad, India, February 19–22.

Patent Curtis and Heafield, 2008. Systems and Methods for Identifying Similar Documents. US Patent 7958136.

Paper and Poster Browne, Wheatley, Welsh, Seibert, Heafield, Rich, and the GALEX Science Team, 2006. RR Lyrae Stars in the Far Ultraviolet: GALEX Observations Compared with Theoretical Predictions. Bulletin of the American Astronomical Society, January.

Article Welsh, Wheatley, Heafield, Seibert, et al., 2005. The GALEX Ultraviolet Variability Catalog. The Astronomical Journal 130, 825–831.

Paper and Poster Welsh, Wheatley, Heafield, Seibert, Browne, and the GALEX Science Team, 2005. The Flaring UV Sky. Bulletin of the American Astronomical Society, January.

Program Committees

European Association for Computational Linguistics	2012
Workshop on Machine Translation	2011
Transactions on Asian Language Information Processing	2011
Machine Translation Journal	2010

Publications are available at <http://kheafield.com/professional/>.