

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

Kenneth Heafield

<r at 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; 3.95/4.0 GPA.

With my adviser Alon Lavie, I work on machine translation system combination. I have performed system combination in the NIST Open MT, DARPA GALE, and Workshop on Machine Translation evaluations. Many sites contribute translations to these evaluations which are of competitive quality but nonetheless differ significantly. My research focuses on combining these translations to produce an improved translation using techniques adapted from statistical machine translation. Evaluations have recently added tracks specifically to evaluate system combination, in which I have shown improvement from 1 to 5 BLEU points over the best individual system. Human evaluation from the Workshop on Machine Translation found my output is best within margin of error for English translations from each of the five other languages.

Bachelor of Science, Caltech

September 2003–March 2007

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

Beyond the required courses, I focused on formal language theory, distributed systems, information theory, and combinatorics. As a student, I:

- Worked for two research projects: Netlab and Galaxy Evolution Explorer, yielding two conference posters and a journal article.
- Did a summer internship in Bangalore at Infosys, yielding a conference paper and patent application.
- Worked for the IT department as a dormitory tech and security tester.
- Represented undergraduates on the Computing Advisory Committee.
- Finished a quarter early.

Skills

Languages

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

Software

Taught Hadoop; Administered Linux, PostgreSQL, and Apache; Used MySQL, Octave, Gnuplot, and GTK

Awards

National Science Foundation Graduate Research Fellowship

2008–

\$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
- Summer Undergraduate Research Fellowship** 2004
\$5,000 for ten weeks of summer research

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.

Conferences

- Presentation** Heafield. CMU-StatXfer Group System Combination. Proc. *NIST Open MT Workshop 2009*, Ottawa, Canada (August 31-September 1, 2009).
- Paper and Poster** Heafield, Hanneman, Lavie. Machine Translation System Combination with Flexible Word Ordering. Proc. *EACL 2009 Fourth Workshop on Statistical Machine Translation*, Athens, Greece (March 30-31, 2009), 56–60.
- Paper** Rama, Sarkar, Heafield. Mining Business Topics in Source Code using Latent Dirichlet Allocation. Proc. *1st India Software Engineering Conference*, Hyderabad, India (Feb 19-22, 2008), 113–120.

Poster Browne, Wheatley, Welsh, Seibert, Heafield, Rich, and the GALEX Science Team. RR Lyrae Stars in the Far Ultraviolet: GALEX Observations Compared with Theoretical Predictions. *Bulletin American Astronomical Society* Poster Sessions 37 (2006).

Poster Welsh, Wheatley, Heafield, Seibert, Browne, and the GALEX Science Team. The Flaring UV Sky. *Bulletin American Astronomical Society* Poster Sessions 36 (2005).

Journals

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

Patents

Application Rama, Heafield, and Sarkar. A Method For Extracting Business Topic From A Source Code. US and Indian applications filed (2008).

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