Jonathan H. Clark

Carnegie Mellon University LTI
Gates Hillman Complex 5407
5000 Forbes Avenue
Pittsburgh, PA 15213

Phone: (412) 254-4566
Office: Gates Hillman Complex 5709
jhclark@cs.cmu.edu
http://www.cs.cmu.edu/~jhclark

Education

Ph.D. in Language Technologies, School of Computer Science **August 2013 *(Expected)***
Carnegie Mellon University, Pittsburgh, PA
Advisor: Alon Lavie
Dissertation: Locally Non-Linear Learning via Feature Induction in Statistical Machine Translation
Application Areas: Machine Learning, Language Modeling, Domain Adaptation

Master of Language Technologies, School of Computer Science **August 2009**
Carnegie Mellon University, Pittsburgh, PA
Advisors: Alon Lavie, Lori Levin, Robert Frederking

Bachelor of Science, Computer Science with minor in Mathematics **May 2007**
Texas Christian University, Fort Worth, TX
Magna Cum Laude
GPA: 3.85 (4.0 in major)

Skills

Languages

- Primary: Scala, Java, C++, Python, Bash
- Secondary: C#, Perl, HTML, CSS

Software

- Java: Ant, Ivy, Hadoop MapReduce, Mallet, JUnit
- C++: Boost, Boost Build, Message Passing Interface (MPI)
- Python: Natural Language Toolkit (NLTK), Scipy, Numpy
- Machine Translation: Moses, Joshua, cdec
- Parsing: Stanford Parser, Berkeley Parser

Experience

Safaba Translation Solutions (Pittsburgh, PA) **June 2010 – Present**
Senior Software Engineer

- Built custom machine translation systems specific to clients' domains
- Automated development workflow to efficiently build and reliably deploy customized translation systems
- Wrote realtime production translation server that scales resource usage to meet current load demands
- Designs implementations of product specifications and mentors junior engineers in their development

Carnegie Mellon University, Language Technologies Institute **August 2007 – Present**
Ph.D. Student

- Implemented "navigator" to elicit dynamically targeted parallel corpora for low-resource languages
- Developed Hadoop MapReduce applications for distributed training of translation and language models
- Wrote stochastic gradient descent (SGD) training distributed over message passing interface (MPI) with L1 regularization for hidden-alignment conditional random field (CRF) with over 30 million features
- Researches application of machine learning techniques to improve optimization in translation systems

Language Computer Corporation (Richardson, TX) **May – July 2006, 2007**
Machine Translation Intern

- Interned with a research-based natural language processing (NLP) company
- Built system to collect comparable/parallel corpora for languages with scarce resources
- Contributed to submission for NIST MT 2006 government evaluation in Arabic and Chinese to English
- Integrated machine translation and question-answering components for CLEF 06 (Cross-Language Question Answering Evaluation) submission

Google Summer of Code **June - August 2005**
Student Developer

- Coded for Gaim, a multi-platform multi-protocol Instant Messaging Tool
- Gained professional experience from the mentoring of Mark Doliner and Gaim developers
- Authored 25-page document that became the basis for implementing OSCAR file transfers in both Kopete in the 2006 Summer of Code and the SHAIM client in fall 2006

Crescent Lab for Intelligent Systems (TCU Dept. of Computer Science) **2005 - 2007**
Undergraduate Research Assistant

- Acted as Lead Student Research Assistant, organizing a team of researchers
- Worked one-on-one with Dr. Charles Hannon, assisting in his research
- Selected by department to represent TCU at the NCUR 2007 conference

Ellis Jewelers, Inc. (North Little Rock, AR) **Summers 1998 – 2005**
System Administrator

- Created custom automatic daily backup system in Java, tailored to store needs
- Designed user interfaces using Access forms and VBA / SQL aimed at ease of use
- Provided software troubleshooting, hardware maintenance, and network administration

ACM Programming Contests **2004 – 2007**
Team Member

- Winner of TCU individual programming contest two consecutive years using Java (judged based on solving the most algorithm-based problems in shortest time)
- Founding member of "/usr/bin/tcu," a campus organization to learn new algorithms
- Selected as part of team for ACM South-Central Regional Programming Competition

Projects

Ducttape – *https://github.com/jhclark/ducttape* **2011 - Present**

- Wrote command line-based hyperworkflow manager as a follow-on to the LoonyBin project (see below)
- Defined formalism for representing workflows that execute many hyperparameter configurations
- Constructed system of tracking all software versions used in a workflow to ensure consistency of results
- Uses: Scala, Bash, Simple Build Tool, Specs Unit Testing

BigFatLM MapReduce Language Model Builder **2010**

- Implemented Hadoop MapReduce training with modified Kneser-Ney smoothing for large language models
- Produced exactly estimated language models on corpora with billions of tokens
- Used: Hadoop, Java 6, Ant

LoonyBin HyperWorkflow Manager – *http://loonybin.sourceforge.net/* **2009 - 2010**

- Wrote Java program to unpack workflows represented as HyperDAGs into bash scripts
- Included web UI to easily monitor status of experiments and system building
- Gained users in both academia and industry
- Used: Java 6, Eclipse, Ant, Ivy

Treegraft SCFG MT Decoder - *http://code.google.com/p/treegraft/* **2008**

- Wrote Java machine translation decoder based on the CMU Statistical Transfer concept including a synchronous context free transducer and a second-stage decoder
- Reviewed literature on core issues in state-of-the-art synchronous parsing technologies
- Used: Java 5, Eclipse, JUnit

Akerblad Sentence Aligner - *http://akerblad.sourceforge.net/* **July 2007**

- Ported LDC's Champollion sentence aligner to Java
- Crafted plugin-based architecture to allow for future research-oriented development by others
- Used: Java 5, Eclipse

Gaim (Pidgin) Instant Messaging Client - *http://www.pidgin.im* **2005**

- Analyzed the behavior of the OSCAR instant messaging protocol and official AIM (AOL Instant Messenger) clients and integrated file transfer capability via proxy server into Gaim (now Pidgin)
- Collaborated with team of developers
- Used: C, GDB (GNU Debugger), CVS, Ethereal Packet Analyzer

Publications

J. Clark, A. Lavie, C. Dyer, " One System, Many Domains: Open-Domain Statistical Machine Translation via Feature Augmentation", *Association for Machine Translation in the Americas (AMTA),* October 2012. San Diego, California.

J. Clark, C. Dyer, A. Lavie, N. Smith, "Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability", *Association for Computational Linguistics (ACL),* July 2011. Portland, Oregon.

C. Dyer, **J. Clark**, A. Lavie, N. Smith, "Unsupervised Word Alignment with Arbitrary Features", *Association for Computational Linguistics (ACL),* July 2011. Portland, Oregon.

C. Dyer, K. Gimpel, **J. Clark**, N. Smith, "The CMU-ARK German-English Translation System", *The Sixth Workshop on Statistical Machine Translation (WMT11) at Empirical Methods in Natural Language Processing (EMNLP),* July 2011. Edinburgh, UK.

G. Hanneman, **J. Clark**, A. Lavie, "Improved Features and Grammar Selection for Syntax-Based MT", *The Fifth Workshop on Statistical Machine Translation (WMT10) at the Association for Computational Linguistics (ACL),* July 2010. Uppsala, Sweden.

J. Clark, J. Weese, B. Ahn, A. Zollmann, Q. Gao, K. Heafield, A. Lavie, "The Machine Translation Toolpack for LoonyBin: Automated Management of Experimental Machine Translation HyperWorkflows", *Prague Bulletin of Mathematical Linguistics (Presented at Fourth Machine Translation Marathon)* January 2010. Dublin, Ireland

J. Clark, A. Lavie, "LoonyBin: Keeping Language Technologists Sane through Automated Management of Experimental (Hyper)Workflows", *The Seventh Language Resources and Evaluation Conference (LREC)*, May 2010. Malta.

G. Hanneman, V. Ambati, **J. Clark**, A. Parlikar, A. Lavie, "An Improved Statistical Transfer System for French–English Machine Translation", *The Fourth Workshop on Statistical Machine Translation (WMT09) at the European Association for Computational Linguistics (EACL),* March 2009. Athens, Greece.

J. Clark , R. Frederking, L. Levin "Inductive Detection of Language Features via Clustering Minimal Pairs: Toward Feature-Rich Grammars in Machine Translation", *The Second Workshop on Syntax and Structure in Translation (SSST) at the Associtation for Computational Linguistics (ACL)*, June 2008. Columbus, Ohio.

J. Clark , R. Frederking, L. Levin "Toward Active Learning in Corpus Creation: Automatic Discovery of Language Features During Elicitation", *The Sixth Language Resources and Evaluation Conference (LREC)*, May 2008. Marrakech, Morocco.

J. Clark, C. Hannon, "A Classifier System for Author Recognition Using Synonym-Based Features", *Sixth Mexican International Conference on Artificial Intelligence*, November 2007.

J. Clark, C. Hannon, "An Algorithm for Identifying Authors Using Synonyms", *ENC 2007*, September 2007.

C. Hannon, **J. Clark**, "A Cognitive-Based Approach to Learning Integrated Language Components", *The Third International Workshop on Natural Language Understanding and Cognitive Science,* May 2006.

M. Bowden, M. Olteanu, P. Suriyentrakorn, **J. Clark**, D. Moldovan, "LCC's PowerAnswer at QA@CLEF 2006," *CLEF 2006 Working Notes,* September 2006.

Posters, Presentations, and Invited Talks

J. Clark "Locally Non-Linear Learning via Feature Induction in Statistical Machine Translation", *Johns Hopkins University Center for Language and Speech Processing Seminar*, October 2012. (Invited Talk)

J. Clark, "Ductape: Automation with HyperWorkflows" , *Carnegie Mellon University Language Technologies Institute Student Research Symposium.* August 2012. (Poster)

J. Clark, "LoonyBin: Automate Your Research", *Carnegie Mellon University Language Technologies Institute Student Research Symposium.* September 2011. (**Best Poster Award**)

J. Clark , J. Gonzalez "Coreference Resolution: Current Trends and Future Directions", *Language and Statistics Literature Review*, Fall 2008.

J. Clark, C. Hannon, "An Algorithmic Approach to Recognizing Authors Using Synonym Weighting", *TCU Student Research Symposium*, April 2007. (**Best Poster Award**, Computer Science Department)

J. Clark, C. Hannon, "A Computational Method for Identifying Authors Using Synonyms", *National Council on Undergraduate Research 2007*, April 2007. (Presentation)

J. Clark, C. Hannon, "Author Attribution via Synonyms", *SIGCSE 2007 Student Research Competition*, March 2007. (Poster)

J. Valentino II, **J. Clark**, "The Cerberus System: Using Artificial Vision for 3D Simulation and Robot Navigation", *TCU Student Research Symposium,* April 2005. (**Best Poster Award,** Computer Science Department)

Teaching

Future Faculty Program – *CMU Eberley Center for Teaching Excellence* **Spring 2013 *(Expected)***

- Studied techniques for effective techniques through seminars
- Discussed and debated teaching techniques with colleagues

Principles of Imperative Computation **Fall 2012**
Teaching Assistant

- Emphasized unit testing and proving code correctness in a C-like language in an early undergraduate course
- Conducted recitations twice a week
- Developed homework content

Algorithms for Natural Language Processing
Teaching Assistant **Fall 2011**

- Taught recitations on subjects including formal language theory proofs and coding NLP applications
- Overhauled homeworks to include coding elements such as the use of OpenFST and implementation of agenda-based parsing algorithms

Grants and Awards

- Supported by Research Assistantship at Carnegie Mellon University (Fall 2007 – Present)
- Awarded $1500 grant from Texas Christian University for Natural Language Research (Spring 2006)
- Mentored research assistants in writing an additional $3000 of successful grant proposals (Spring 2006)
- Earned Dan Drew Scholarship from Upsilon Pi Epsilon national honor fraternity for the computing sciences (Spring 2006)
- Funded by Texas Christian University Dean's Scholarship for duration of undergraduate career
- Named Senior Scholar in Computer Science for Texas Christian University Class of 2007 (Spring 2007)

Service

Reviewing

- Empirical Methods in Natural Language Processing (EMNLP) 2011
- European Association for Machine Translation (EAMT) 2011
- Workshop on Statistical Machine Translation (WMT) 2010

Organization

- Carnegie Mellon University Language Technologies Institute Student Research Symposium (SRS) 2012

Relevant Graduate Coursework

Machine Learning

- 10-701: Machine Learning: Mathematical theory and MATLAB implementations of core machine learning including Naïve Bayes, perceptron, boosting, neural networks, Bayesian Networks, and reinforcement learning
- 11-765: Active Learning Seminar (Audit): Presented papers on current methods including uncertainty sampling, density sampling, and combined strategies

Natural Language Processing

- 11-761: Language and Statistics I: Basic NLP methods such as maximum entropy training, expectation maximization, Hidden Markov Models, and smoothing techniques
- 11-762: Language and Statistics II: Modern NLP methods such as conditional random fields, non-parametric Bayesian Inference, Gibbs Sampling, the Inside-outside algorithm, and semi-supervised learning
- 11-731: Machine Translation: Current methods for statistical MT
- 11-713: Advanced Natural Language Processing Seminar
- 11-712: NLP Lab: Coded "Treegraft," a statistical SCFG decoder (see below)
- 11-734: Advanced Machine Translation Seminar (Audit): Presented papers on current MT research
- 11-711: Algorithms for Natural Language Processing: Presented papers and dissertations on current methods for NLP

Linguistics

- 11-722: Grammar Formalisms: Studied and wrote formal grammars in Head-driven Phrase Structure Grammar, Lexical Functional Grammar, and Combinatory Categorial Grammar
- 11-721: Grammars and Lexicons: Included language typology and its impact on NLP