Keith D. Stevens

Contact Information

Institute for Scientific Computing Research
Lawrence Livermore National Lab 7000 East Ave,
Livermore, CA 94551

Mobile: +1-805-433-2295
fozziethebeat@gmail.com
http://fozziethebeat.github.com

Objective

Placement in a development and/or research position centered around distributed semantics and unsupervised learning, along with applying current theories on these topics to web search or other human oriented data exploration tasks.

Research Focus

My research focuses on the unsupervised acquisition of word senses and the automated evaluation of related models. Word sense acquisition, or Word Sense Induction (WSI), attempts to use contextual information to learn distinct word senses as they appear in large corpora. My PhD centers around using Ensemble Clustering to improve Word Sense Induction by combining multiple models into a single system. Using this approach, I not only find ways to find some improvement over existing approaches, I use the added knowledge created by multiple models to survey existing state of the art methods and evaluate the effectiveness of current evaluation techniques. Beyond Word Sense Induction, I am interested in automated approaches for navigating large document sets based on semantic knowledge and interests.

Education

University of California, Los Angeles, Los Angeles, California USA

Ph.D., Computer Science, In Progress

- Thesis Topic: *Ensemble Clustering and Word Sense Induction: when is it right for Word sense Induction?*
- Adviser: Professor Michael G. Dyer
- Area of Study: Computational Linguistics

M.S., Computer Science, October 2011

- Thesis Topic: *Extending Word Sense Induction with waiting and depending*
- Adviser: Professor Michael G. Dyer
- Area of Study: Computational Linguistics

B.S., Computer Science, December 2007

- Computer Science
- Minor in Mathematics

Research Experience

Lawrence Livermore National Lab, Livermore, California USA

Ensemble Clustering and Word Sense Induction
June 2011 - *present*
Advisor: *David Buttler*

Ensemble Clustering is a new approach that combines together multiple clustering models into a single ensemble approach that closely matches ensemble methods in supervised learning. I am applying these Ensemble Clustering approaches to the task of Word Sense Induction in order to both leverage gains in existing state of the art methods and gain a better understanding of why existing models fail on current shared tasks. Using this approach, I've found that existing models fail to perform adequately on shared tasks due to a general lack of accuracy and developed new semantic measures. Using these new semantic measures, I intend to show that current tasks fail to measure the value of Word Sense Induction models appropriately and need to be significantly redesigned.

and Topic Model Evaluation
June 2011 - *present*
Advisor: *David Buttler*

I evaluated three topic models with two recently developed coherence metrics that have been closely associated with human judgements. Our current results indicate that the coherence metrics, originally designed for Latent Dirichlet Allocation, can be applied to other techniques such as Non-negative Matrix Factorization and Singular Value Decomposition, as seen in Latent Semantic Analysis). Furthermore, we are evaluating new aggregate metrics for comparing entire models, rather than just individual topics.

Based on the above research, I am beginning to investigate the applicability of these coherence metrics as an intrinsic evaluation of Word Sense Induction (WSI) models. Current evaluation strategies for WSI focus on costly human judgements and do not provide a method for improving the models. Along these lines, I am also Consensus Clustering as an evaluation of the feature space used in WSI models and the stability of clustering models used to differentiate word senses.

Ontology Expansion
June 2010 - September 2010
Advisor: *David Buttler*

With David Buttler, I have investigated methods of automatically enhancing existing concept hierarchies for the purpose of organizing unstructured text documents. This work is in coordination with my work in Distributional Semantics and will soon be used by several information retrieval tasks at the national lab.

University of California, Los Angeles, Los Angeles, California USA

Distributional Semantics
September 2008 - June 2011
Advisor: *Michael G. Dyer*

Distributional semantics assumes that the semantics of words can be defined by the contexts in which they occur, an idea proposed by Firth in 1957. The learned semantics have been shown to approximate semantic judgements made by humans on psychological and standardized tests. My research in this field has three main goals.

First, models for distributional semantics have been developed in a variety of fields, such as Computational Linguistics, Cognitive Science, and Artificial Intelligence. In order to permit accurate evaluation of distinct models, a common framework is needed for their implementation and evaluation. To this end, I have collaborated with David Jurgens to develop the S-Space Package, a java based framework for implementing, testing, and extending distributional models.

Second, current distributional models are incapable of distinguishing between distinct meanings for a single word, for example, cat can refer to feline pets, large felines found in the wild, or a brand of construction vehicles. To learn these differences, I combine distributional models with clustering techniques. Initial research in this field, Word Sense Induction, has already been shown to perform well according to

one measure in a coordinated semantic task, SemEval, and to be applicable towards detecting the occurrence of news worthy events.

Lastly, current distributional models lack organization. Word semantics are unorganized and disconnected from each other. In contrast, concept hierarchies provide a rich set of relationships between the semantics associated with many words. I investigate methods of combining distributional models and Word Sense Induction models for the purpose of automatically building concept hierarchies. For this, I focus on two approaches: extending existing hierarchies by enhancing current semi-supervised techniques and inferring hierarchies automatically from word semantics learned through only distributional methods.

Publications

Referred

Keith Stevens, "Evaluating Unsupervised Ensembles when applied to Word Sense Induction." to appear in *Association of Computational Linguistics Student Research Workshop 2012*.

Keith Stevens, Terry Huang and David Buttler, "The C-Cat Wordnet Package: An Open Source Package for modifying and applying Wordnet." *Systems Demonstration of the Global Wordnet Conference 2012*, 2012.

David Jurgens and Keith Stevens, "Measuring the Impact of Sense Similarity on Word Sense Induction." *Proceedings of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, 2011. ACL.

David Jurgens and Keith Stevens. "Capturing Nonlinear Structure in Word Spaces Through Dimensionality Reduction." *Proceedings of the ACL 2010 Workshop on GEometrical Models of Natural Language Semantics*, 2010.

David Jurgens and Keith Stevens. "HERMIT: Flexible Clustering for the SemEval-2 WSI Task." *Proceedings of the ACL 2010 SemEval-2010 Workshop*, 2010.

David Jurgens and Keith Stevens. "The S-Space Package: An Open-Source Framework for Word Space Algorithms." *Proceedings of the ACL 2010 System Demonstrations*, 2010.

David Jurgens and Keith Stevens. "Event Detection in Blogs using Temporal Random Indexing." *In Proceedings of the International Workshop on Events on Emerging Text Types* (eETTs), pages 21-28, 2009.

Manuscripts

David Jurgens and Keith Stevens. "Word Ordering with Reduced Dimensionality for Sense Induction." Technical Report 090020, University of California Los Angeles, 2009.

Keith Stevens. "Extending Word Sense Induction with waiting and depending." Masters Thesis, University of California Los Angeles, 2010.

Teaching Experience

University of California, Los Angeles, Los Angeles, California USA

Computer Science 32 - Introduction to Computer Science II
Teaching Assistant, *Winter 2009 Spring 2009*

This course introduces students to basic object oriented design a fundamental data structures. Topics include C++ classes, C++ templates, pointer management, trees and traversal methods, inheritance, and recursion. As the teaching assistant, I was responsible for teaching a weekly 2 hour discussion section that reviewed class lectures and broke down class assignments.

Computer Science 35L - Software Construction Laboratory
Teaching Assistant, *Fall 2008 Fall 2009*

This course introduces students to fundamental concepts for developing software in a unix environment. Topics include Bash and Python scripting, The GNU debugger, basic C programming, version control, network security, and basic operating system tools. As the teaching assistant, I was responsible for all lectures and lab assignments.

Computer Science 111 - Introduction to Operating System Principles
Teaching Associate, *Spring 2010 Fall 2010*

This course introduces students to the basic theories behind modern operating systems, with a focus on theories behind the Linux Kernel. Topics include concurrency, device management, file systems, memory management, process and thread management, Remote Procedure Calls, scheduling, security and protection, and shell design. As the teaching assistant, I applied theories taught in class to a six lab assignments that covered virtual memory, process management, kernel modules, shell design, file systems, and remote applications.

Computer Science 132 - Modern Compiler Construction
Teaching Associate, *Winter 2011*

This course introduces students to the theories and practices behind compiler design. Students learn topics focusing on parsing, type checking, register allocation, and virtual machines. As the teaching assistant, I applied theories taught in lectures to practical examples of compilers and assisted students with lab assignments, along with grading of all lab assignments.

Engineering 183 - Engineering and Society
Teaching Associate, *Winter 2011*

This course introduces students to professional and ethical considerations experienced in engineering positions. Emphasis is placed on the impact of technology on society and the development of moral and ethical values. This course also serves as the technical writing course for the School of Engineering. Lectures cover ethical topics, while teaching assistant-led sections cover technical writing.

I taught a weekly three-hour discussion on technical writing and ethics. As the teaching assistant, I covered all writing requirements for students and provided a series of writing exercises designed to improve their ability to concisely describe, evaluate, and solve ethical dilemmas in Engineering.

Professional Experience

Lawrence Livermore National Lab, Livermore, CA

Student Research Intern, **June 2010 to September 2010**
Mentor: *David Buttler*

- Implemented an existing taxonomy expansion algorithm for a highly distributed framework based on Hadoop and HBase
- Implemented several methods for condensing an existing taxonomy
- Evaluateedseveral methods for integrating taxonomy induction and condensation techniques
- Evaluated several methods for enhancing taxonomy expansion and condensation methods such that the resulting taxonomy is customized for a particular domain of knowledge
- Developed a new, java based interface for the widely used WordNet taxonomy

Google, Inc., Kirkland, WA

Software Engineering Intern, **June 2009 to September 2009**
Mentor: *Zhen Lin*

- Enhanced a document selection system with the addition of semantic analysis. This analysis provided a more detailed representation of the document, as it changed over several points in time. This analysis allowed the application to inform users of more relevant aspects of the page.
- Enhanced crawling framework for the document selection system, improving savings by up to 50% usage of external resources.
- Restructured document selection system to be more scalable and abide by internal policies.

Software Engineering Intern, **January 2008 to September 2008**
Mentor: *Hizakazu Igarashi*

- Built a highly scalable, stateless server for responding to a wide variety, and high volume, of requests.
- Developed several plugins for the designed server allowing access to a wide variety of data sources.
- Improved serving time of the server by addition of caching responses in memory
- Improved response time of a query formulation system with the use of a Trie, and integration as a plugin for the server.
- Heavily unittested, document, and debugged same server.

Networked and Embedded Systems Laboratory, Los Angeles, CA

Student Research Intern, **March 2007 to December 2007**
Mentor: *Professor Mani Srivastava*

- Designed multiple modules for use on a Micaz/Mica2 sensor board using the SOS embedded operating system, which is based on a Linux architecture.
- Reduced the transmission rate of nodes by defining a time series representing sensor values.
- Designed a unit test framework for the SOS operating system.

Open Source Projects

The S-Space Package

The S-Space package is an open source package written in Java for building and evaluating word space algorithms learned from word distributions. The package includes reference implementations of frequently cited algorithms, specialized data structures for natural language processing, and multi- threaded matrix implementations for concurrent algorithms.

It also now includes a framework for building Word Sense Induction algorithms and a growing library of clustering algorithms.

I work on this this project as a primary contributor with David Jurgens. This package is being developed as a part of my research, and makes all of the software used for my research freely available. The project is available on GitHub at www.github.com/fozziethebeat/S-Space

The C-Cat Package

The C-Cat package provides three key features for processing large corpora over the Hadoop MapReduce framework : text processing utilies such as parsing, tokenization, and sentence detection; a new WordNet API that facilitates direct modification of the lexical ontology; and new APIS and MapReduce codes to automatically extend or reduce an ontology based on distributional information.

This project was developed while I worked for Lawrence Livermore National Lab. It also serves as a staging ground for MapReduce based word space implementations based on those in the S-Space package. I am the primary contributor. The project is available on GitHub at www.github.com/fozziethebeat/S-Space

Service

The Computer Science Graduate Student Committe

Contributing member of a UCLA graduate student group dedicated to improving the academic and social lives of computer science graduate students. My contributions include

- Organizing and advertising for a presentation competition called *So You Think You Can present*: **$CONTEST** aimed at having graduate students talk about exciting research ideas in front of large audiences.
- Organizing for bi-annual department picnics.
- Organizing for weekly tea-times.
- Mentoring incoming Masters and PhD students.

The Engineering Graduate Student Association

Board member of a UCLA Engineering Graduate Student Association, a student group dedicated to improving the academic and social lives of engineering graduate students. I served as the *Sustainability Chair*, my contributions include

- Attending meetings with other student groups to discuss current sustainability practices at UCLA.
- Organizing and managing a quarterly Engineering Social.
- Organizing information sessions.

UAW Local 2864

Head Steward of the UAW Local 2865, the union representing Teaching Assistants, Readers and Graders for all schools in the University of California system. My responsibilities include

- Organizing an election campaign.
- Managing the UCLA unit's blog.

- Representing engineers in quarterly union wide meetings.
- Organizing and presenting for information sessions to raise awareness and interest in the union.

The University Buddhist Association

The UBA provides an open non-sectarian group for students that are practicing Buddhism or interested in learning more about the religion. I served as a leading member to organize socials, field trips to monasteries and temples, and connect with other Buddhist groups in Southern California.

Mentoring

Sky Lin, Grace Park, and Alex Nau, 2009
I worked closely with three students as a part of the CS 199 Course: Directed Research in Computer Science. As their mentor, I helped select and conduct several concen- trated research tasks in Distributional Semantics. Their work focused on implementing and evaluating useful distributional models and has been incorporated into the S-Space Package

Alexander Honda, 2010
I worked with Alexander on two tasks: examining feature importance in word-space models and studying topic modelling. This work was done as a part of the CS 199 Course: Directed Research in Computer Science.

David Cohen, 2010
I worked with David on extending lexical ontologies with word-space models. This work was done as a part of the CS 199 Course: Directed Research in Computer Science.

Terry Huang, 2011
I mentored Terry Huang in the Winter and Spring quarters of 2011 to build an open source implementation of the Castanet algorithm, an automated method of learning hierarchical facets for a document set based on distributional information and the WordNet lexical ontology. Since the summer, we have been incorporating it into the C-Cat package and co-authored a recent paper submitted to the Global Wordnet Conference of 2012.

Patents

"Semantic Document Analysis," with Zhen Lin. FR16113-2118P01; GP-2700-00-PR

"Efficient scheduling of webpage crawls," with Zhen Lin. FR16113-2119P01; GP-269900-PR

References

Available upon request