Erin M. Tavano, Ph.D.

817 Concord St.
Carlisle, MA 01741
Phone: (978) 893-8664
Email: emtavano@gmail.com
http://www.lingchat.net/erin

SKILLS

Natural language processing
Computational linguistics
Corpus linguistics
Theoretical linguistics
Machine learning
Software development
Advanced research and writing
Psycholinguistics
Experiment design
Large-scale data

EDUCATION

Ph.D. in Linguistics
University of Southern California, Los Angeles, CA, 2010
Concentration: Psycholinguistics, computational linguistics, pragmatics, semantics

B.S. in Computer Science
Northeastern University, Boston, MA, 1998

TECHNICAL KNOWLEDGE

Operating systems:	Windows, Linux, Mac OS X
Programming languages:	Java, Python, SQL; Less recently: Perl, C++, Visual Basic / Visual C#
Web development technologies:	Less recently: Google Web Toolkit, PHP, ASP.NET, HTML, XML, CSS, Apache, Microsoft IIS, Jetty / Tomcat
Databases:	Experience in both RDBMS and NoSQL: MySQL, HBase, MongoDB, Oracle, MS SQL Server, Cassandra
Other:	Hadoop and MapReduce; Amazon EC2 and Elastic MapReduce

EMPLOYMENT EXPERIENCE

NLP Data Scientist
January 2017 - Present

Clinical NLP Specialist
September 2015 – January 2017
Linguamatics
Westborough, MA

- As NLP Data Scientist, I work with senior management to create new products, analyze current customer use of the current products and data, and on company partnerships. Currently working on creating **machine learning** models to predict medical outcomes based on information in **unstructured medical record text,** similar to IBM Watson. (Using **Python, SVM, Scikit-learn**)
- Wrote, from scratch, a re-implementation of the Linguamatics I2E text mining software software in **Java** and **HBase**, using **Amazon Elastic MapReduce**, in order to help the company expand into the **Hadoop** market.
- As NLP Specialist, primarily worked as a consultant for I2E, in a customer-facing role with varied technical needs.This included training and professional services, that is, using I2E to develop **information-extraction** queries on medical records.
- Used **Python** to clean and perform other preprocessing of data, develop ontologies, and conduct evaluations.

NLP Engineer/ Senior Software Engineer, Humedica
Boston, MA
July 2014 – September 2015

- Implemented software to perform **statistical text analysis**, **information extraction** and **deidentification** on 2.2 billion medical records, using Java, Python and C++. Wrote and documented code daily, mostly Java.
- Responsible for “monthly run,” the complete re-parsing of all records every month. Created and ran procedures for adding new records to NLP processing system.
- Implemented improvements to parsing system that reduced parse time by 40%. Efficiently indexed large text files of aggregated medical records in an **Oracle database**.
- Used complex **context-free grammar (CFG) system with feature structures** to parse medical records. Wrote proposals, ontologies, and grammars for new modules (alcohol consumption, smoking, e-cigarette use, EKG measures)
- Using **Java**, **Google Web Toolkit (GWT),** and **MySQL,** designed, implemented and documented a web-based system for labeling gold-standard data, producing industry-standard scores and creating graphical diffs of system output files.

- Wrote deidentification software to remove protected health information (PHI) while maintaining readability and preserving non-PHI text for use in NLP.
- Created evaluation standards and managed annotation projects for consistent evaluation of current data quality. Created analytics to identify sets of records whose data may be parsed incorrectly.

NLP Research Scientist, Leidos (formerly known as SAIC)
Reston, VA
August 2012 – July 2014

- For DARPA Machine Reading, design and implement **Natural Language Processing (NLP)** algorithms for **relationship extraction** and **entity extraction**, using linguistics-informed **machine learning techniques,** primarily support vector machines (SVM) and conditional random fields (CRF).
- For DARPA DEFT, implement original NLP approaches to non-prose documents, such as resumes, which are generally not well handled by standard extraction techniques.
- Research, evaluate and implement NLP applications for a government customer dealing with large amounts of classified text. Because most NLP models are trained on unclassified text, they must be adapted; labeling classified data is not usually an option. Applications include **entity and relationship extraction**, **document classification / summarization**, **entity disambiguation**. A particularly large project was **document clustering**, which had to be performed incrementally as new documents were received, and which was based on the customer's ontologies.
- All development in Java, with code written on a daily basis.
- Also maintain oversight of the state of the customer's **NLP pipeline**, which integrates code from multiple contributors. Supervise and set goals for other contractors providing NLP software, including evaluation and data labeling (annotation) efforts.
- Provide technical direction and leadership for NLP projects for government customers. Write and review technical proposals for DARPA, IARPA and other research agencies.

NLP Scientist, Orbis Technologies
Annapolis, MD
October 2011 – August 2012

- Research, design and develop **Natural Language Processing (NLP)** components of Orbis' Cloud Text Analytics (CTA) software, which runs on **Hadoop** clusters.
- Adapt CTA for use on real-world social media problems, such as finding the person who is the originator of a piece of information.
- Develop and manage procedures for improving **entity extraction** models, including **human annotation** of data.
- **Manage software developers** in NLP efforts from evaluation scripts, scoring and corpus analysis through implementation of new NLP components.

- Implemented algorithm improvements to entity extraction and sentence detection software for **large-scale data**.
- Designed and implemented original **relationship extraction** software based on recursive comparisons of a reference constituent parse tree against a test tree.
- Consult on NLP problems and proposals.
- Work with clients to adapt CTA to datasets of many different types.

Senior Consultant, Invertix Corporation
Alexandria, VA
January 2011 – October 2011

- Research and develop **Natural Language Processing (NLP)** and **text analytics** solutions for **large-scale data problems**, including the evaluation of **entity extraction** software and integration of multiple extractors in a **Hadoop MapReduce** framework.
- Lead a team to migrate a **Cassandra**-based store of processed text to **Cloudbase**.
- Discover relationships in unstructured text, stored in a Cassandra database, and assist in the development of web-based display widgets. This requires the integration of multiple **ReST** and other data services.
- Research and advise on computational linguistics / NLP projects and contribute to related proposals.

Programmer analyst, Linguistic Data Consortium
Philadelphia, PA
June 2010 – January 2011

- Worked on the evaluation team of a DARPA study, Machine Reading, that evaluated the ability of machines to find information in unstructured text.
- Designed and developed a web-based set of tools for users from the general public to annotate text according to a prescribed format, using **Google Web Toolkit, Java, PHP, MySQL, XML, and Lucene SOLR** search engine.
- Helped create the queries that the machines would have to answer by designing and implementing a web-based tool for other members of the evaluation team (from SAIC) to design queries for the evaluation.
- Researched crowd-sourced linguistic annotation techniques and designed a system that would allow **Amazon Mechanical Turk** users to annotate short YouTube videos. For instance, one person would describe the main activity in the video and another person would verify that it had that topic. This project integrated several different APIs, such as those from Amazon and YouTube.

Summer Research Assistant, USC Information Sciences Institute
Marina Del Rey, CA
Summer, 2007 and Summer, 2008
Supervisor: Dr. Jihie Kim, Dr. Eduard Hovy

- Worked as a research assistant for the ISI PedDiscourse (Pedagogical Discourse Analysis) project. The goal of the project was to develop a **question answering system** (PedaBot) to help students who were participating on an online course message board to find answers to their questions about the class.
- Designed a new annotation scheme to describe the activities of students, and helped other researchers describe and implement annotation strategies.
- Using **Perl** and techniques such as **Latent Semantic Analysis** and **support vector machines**, analyzed messages to identify those that were similar, and which might have similar answers. Also conducted **sentiment analysis** on messages, e.g. to detect stress or frustration on the part of the student. This work produced several conference papers and a working system that is used by USC professors.
- Created a **Java** application prototype for extracting relevant answers to questions from tagged text using regular expressions.

Graduate Research Assistant, USC Institute for Creative Technologies
Marina Del Rey, CA
Summer, 2005
Supervisor: Dr. Andrew Gordon

- As a graduate research assistant on the ICT **Knowledge Representation** team, conducted research investigating linguistic and **evaluation** issues in **commonsense psychology** (how people believe their minds work, and how that knowledge can be formalized towards solutions in **artificial intelligence** problems) and knowledge representation in first-order logic.
- Created a linguistic resource for commonsense psychology terms.
- **Conducted and wrote original research into how people express their own patterns of thinking and reasoning, based on** Protocol Analysis **techniques.**

Software development consultant
Los Angeles, CA
January 2004 – August 2009

- As a freelance consultant, worked on many different projects, primarily **ASP.NET or VB.NET / MySQL / MS SQL server / MS Access** based websites and reporting software.
- A multi-year project in this period was an order management and **reporting system** for a document copying company CopyPage, which replaced the client's existing paper-based system, becoming widely used in client's multiple locations.

Lead Software Developer, Softman
Marina Del Rey, CA
March 2002 – September 2004

- Lead developer on the popular software retail website BuyCheapSoftware.com, which sells retail boxed software. The site was designed using **ASP**, later rewritten with **ASP.NET**, and **MS SQL Server**, and was visited by 500,000 users per month. Maintained the site, addressing scalability issues, generating reports, and developing and documenting new features on the administrative and front-end sides.
- Developed original software project that offered a customizable, partially open-source shopping cart system written in ASP.NET and **C#**, license-able on a monthly basis from Softman.
- Responsible for all **project management** tasks including management and hiring of other developers.

Software Engineer, Radio Free Virgin
Los Angeles, CA
January 2001 – October 2001

- Worked with a team of software engineers for an online radio station created by the Virgin Entertainment Group. Radio Free Virgin streamed songs and advertising on programmed stations, and offered paid and free membership programs for listeners.
- Using primarily **Visual C++**, wrote components for programming the stations and serving the audio streams, and integrated them into a web user interface.
- Researched and implemented new features, wrote specifications and design documents.

Junior Software Engineer, Turning Point Software (later Metamor)
Newton, MA
September 1998 – December 2000

- Worked on a variety of projects doing primarily Windows development using Visual C++ and Microsoft Foundation Classes (MFC). Primary projects were a grocery delivery service client/server application, and a computer-connected whiteboard.
- Also worked on HTML / ASP / MS SQL Server websites, including retailers for inkjet cartridges and insurance.

Web Developer Intern, PlanetAll
Cambridge, MA
1997-1998

- Using HTML, ASP and MS SQL, assisted in the development of PlanetAll, an early social networking website that allowed users to connect with each others' contacts via an address book, and coordinate meetings according to users' travel plans.