Research Statement
John Lafferty

Introduction

My career goal is to produce outstanding research that contributes in fundamental ways to the advancement of computer science, language technologies, and the interface of computer science and statistics. As an integral part of this goal, I am committed to encouraging the next generation of researchers through teaching, mentoring and collaboration with students.

The broad direction of my research is the development of theoretically motivated methods for addressing practical problems in computer science, typically using ideas and techniques from statistics, probability, and information theory. The applications that I am most interested in advancing are text and natural language processing, information retrieval, and other areas of language technologies.

When I began working on statistical language processing at the IBM Watson Research Center in 1988, the statistical approach was viewed as heretical, radical, and destined for failure. Today, statistical methods are seen as fundamental and essential in most areas of artificial intelligence and many other computational sciences, such as computational biology. However, even while the use of probabilistic methods represents a paradigm shift that I believe is here to stay, it is clear that our current approaches and methodology are inadequate to solve most of the hard problems. The objective of my research is to achieve fundamental advances in these areas.

Research Areas and Recent Accomplishments

My research spans a number of areas, including statistical machine learning, text and natural language processing, information retrieval, information theory, and computational group theory. The following sections sketch some of the work in these areas that I have recently been involved in with colleagues and students.

Natural Language Processing

At the IBM Watson Research Center and during my career at CMU, I have researched statistical methods for many problems in natural language processing, including the development of generative models for parsing, non-parametric decision tree methods applied to discriminative models for tagging and parsing, language modeling, and text segmentation. I was a member of Fred Jelinek’s department at IBM, initially working in Peter Brown’s group on the statistical approach to machine translation based on the source-channel paradigm from information theory. This was one of the first large-scale statistical
machine learning projects to successfully demonstrate the power of statistical methods in an area that had been traditionally grounded in hand-built, knowledge-engineered systems.

Experience in statistical machine translation has been very important in my development as a researcher. The problem is fiendishly difficult, and humbling to work on. It makes clear that our current statistical methodology is inadequate. At the same time, working on such a complex problem has taught me the value of rigorous and mathematically clean methods for text problems, and had instilled in me an aesthetic for good modeling practice, even within large systems.

Professionally, I have served on the editorial board for *Computational Linguistics*, the leading journal in the field, and have served on program committees for the major conferences. Lillian Lee (Cornell University) and I presented an invited tutorial at AAAI in 1998 on statistical methods in natural language processing; the tutorial was very popular and we were asked to present it again at AAAI'99. I have also recently presented an invited tutorial on language modeling at NIPS.


**Statistics and Machine Learning**

The study of learning algorithms and statistical methods for their own sake is a major component of my research program. I have been particularly interested in statistical machine learning methods based on the use of random fields and exponential models, sometimes also
referred to as maximum entropy methods. This class of models is widely used for spatial
statistics problems, physics, computer vision, and other areas of computational science. I
have been interested in these methods for a wide range of text processing problems. Most
recently, in work with Fernando Pereira and Andrew McCallum at WhizBang Labs, we
have introduced a modeling framework based on conditional random fields that offers many
advantages over more commonly used methods such as hidden Markov models. Researchers
in computational biology have been excited to learn about this modeling technique, as they
have encountered many of the same issues that make this class of models attractive for
information extraction from text.

In trying to understand the proper abstract setting for such models, Stephen Della Pietra,
Vincent Della Pietra, and I introduced techniques based on a class of generalized distances
called Bregman distances. (The magical properties of Bregman distances were realized in-
dependently in the machine learning community by Manfred Warmuth at UC Santa Cruz.)
Recently Michael Collins, Rob Schapire, and Yoram Singer (AT&T Labs and Hebrew Uni-
versity) have used our work on Bregman distances to prove new properties of boosting
algorithms. Most recently, I have used duality and information geometry to show (in a
sense that can be made precise) that boosting and maximum likelihood for exponential
models optimize the same objective function, subject to the same feature constraints; the
only difference is that the latter uses models that are constrained to be normalized.

In current work with Larry Wasserman (CMU Statistics Department), I am investigating
Monte Carlo approximations to iterative algorithms. Our collaboration was initially fo-
cused on a rather specialized, but important problem in Bayesian statistics: calculating
reference priors for general parametric families. It now appears that our approach to the
analysis of iterative Markov chain Monte Carlo applies to an extremely broad and im-
portant collection of problems, including maximum likelihood estimation for exponential
models, multicommodity flow problems, and state space model approximations using “par-
ticle filters.” For some problems our results may give the first theoretical analysis of a class
of sampling techniques that are widely used in practice. Collaboration between statistics
and computer science is fostered by the Center for Automated Learning and Discovery
(CALD), which makes CMU an attractive place to work in this area.

Professionally, I am currently organizing an invited session on statistical methods for text
at the Interface of Computing Statistics (Interface’01). I serve as an action editor for the
Journal of Machine Learning Research, as well as the journal Machine Learning. This
summer I will present an invited tutorial on exponential models for probabilistic learning
at the UAI (Uncertainty in Artificial Intelligence) conference.

- J. Lafferty, F. Pereira, and A. McCallum, “Conditional random fields: Probabilistic
  models for segmenting and labeling sequence data,” In International Conference on


**Information Retrieval**

The Internet boom has brought the importance of information retrieval to the forefront, and introduced a wide variety of new problems. I have recently become interested in some new statistical approaches to information retrieval using language modeling methods.

In 1998, Bruce Croft and his student Jay Ponte at the University of Massachusetts first proposed a very simple retrieval model using a kind of statistical language model. It is a natural and powerful idea, and the fact that it came as late as 1998 can only be attributed to “sociological” reasons. Adam Berger and I formulated an extension of the Ponte-Croft model using ideas from statistical machine translation. This line of work was developed much further in Berger’s Ph.D. thesis. Recently we have begun a collaboration with the UMass group, under a new project that is funded by ARDA, and have organized a workshop on language modeling and information retrieval. As part of the Lemur project we are developing a software toolkit to support research in language modeling for IR and other applications.

In recent work under this project with Chengxiang Zhai, we have developed a new formal framework for the language modeling approach to IR based on Bayesian decision theory, have shown a simple connection between the standard probabilistic IR model based on document generation and the LM approach, and have developed a technique based on Markov chains that addresses many of the shortcomings of the translation models developed with Berger, with very good empirical results on standard TREC benchmarks. In addition, we have recently carried out a study of smoothing methods for language modeling in IR. The language modeling approach is natural and effective—I anticipate that it will become the standard approach to text-based IR that is taught in future courses and textbooks.

Professionally, I have served as an area coordinator (with Keith van Rijsbergen) for formal models and language models at this year’s SIGIR conference. I have consulted in the past for Lycos and Google to develop language modeling techniques to model users and query logs. I view this as an important complement to my academic research in this area, both
because of the data available and the perspective that such companies have on search and retrieval problems.


### INFORMATION THEORY

Much of my work has benefitted from the ideas and perspective of information theory, which provides a kind of intellectual glue between various aspects of my research. Time and again, I have found that thinking in terms of information theory yields insights and a fresh look at problems that complement traditional statistical and computer science approaches. For example, my recent collaboration with Larry Wasserman started with the insight that the Blahut-Arimoto algorithm for computing channel capacity can be applied to calculate reference priors in Bayesian analysis. As another example, my work with Adam Berger on translation models for information retrieval was inspired by the noisy channel paradigm that is so common in information theory. The fact that this field has been such a source of ideas and inspiration has led me to investigate problems in information theory for their own sake.

The field of error-correcting codes has recently been turned upside down by the rediscovery of very simple randomized constructions of sparse codes that are efficiently decoded near channel capacity using the technique of belief propagation that is familiar to AI researchers. While all of this work uses random code constructions, I have recently shown with Dan Rockmore how explicit families of algebraic expander graphs can be used to construct codes with very similar properties. Our results are among the first to show that codes
from explicit, algebraically defined graphs can outperform random constructions for long block lengths. In our 1997 STOC paper, Rockmore and I also presented an algorithm based on Fourier analysis for non-abelian groups that gave the first sub-quadratic algorithm for encoding low-density codes, solving an open problem posed by Sipser and Spielman.

With Alexander Vardy (University of California at San Diego) I established a close connection between the minimal trellis, a graphical structure that is central to the theory and practice of error-correcting codes, and ordered binary decision diagrams (BDDs) used in computer verification. Using BDDs and code trellises, I have recently developed a decoding technique that I call “projection decoding.” The idea is to decode by taking random projections to form a graph that “covers” the code, and to then use iterative decoding algorithms on the resulting graph. Together with Yair Weiss at Berkeley, I am currently investigating related techniques for constraint satisfaction.

As professional activities in this area, I have served as referee for the IEEE Transactions on Information Theory, as well as a reviewer for the annual information theory conference. I was an invited speaker at the 1999 Workshop on Codes, Systems, and Graphical Models at the Institute for Mathematics and its Applications, which brought together researchers from computer engineering, coding theory, systems theory, artificial intelligence, and statistics (http://www.ima.umn.edu/csg).


**Computational Group Theory**

The fast Fourier transform (FFT) is one of the most important families of algorithms in computational science. In work with Dan Rockmore, I have been studying the FFT for various non-abelian groups, together with the connections to expander graphs and spectral graph theory, topics that are important to many areas of theoretical computer science.

There are several ways to approach the FFT. In the group-theoretic approach, the goal is to obtain an efficient expansion of functions defined on a finite group in terms of a set of irreducible representations of the group. We have developed an FFT for the group of invertible $2 \times 2$ matrices over the finite field of order $q$ that has complexity $O(q^4 \log q)$ in a
model of computation where each complex add or multiply has unit cost. (It is conjectured that an algorithm with complexity $O(q^3 \log^c q)$ should be attainable.) This algorithm was used in our work on coding theory discussed above.

Expanders built on Cayley graphs are an important tool, and there are many promising directions to explore here. For example, the connection recently made between the zig-zag construction of expanders and Cayley graphs for semi-direct product groups suggests many interesting new classes of error-correcting codes.

We have developed software that implements our algorithms over finite fields, and have used this software to conduct extensive experiments, in both spectral graph theory and coding theory.


**The Future**

Recent years have seen a confluence of some of the core research issues in natural language processing, information retrieval, statistics, and machine learning. Such issues as generative versus discriminative modeling, the use of unlabeled data, Bayesian methods, feature selection algorithms, approximation techniques, the use of expert knowledge, and coding-theoretic methods will continue to be important in my research agenda. There is tremendous opportunity for new ideas. I plan to take a more high-risk, high-payoff approach to selected research projects in these directions.