An Introduction to Conditional Random Fields

An Introduction to Conditional Random Fields

Charles Sutton

University of Edinburgh
Edinburgh, EH8 9AB
UK
csutton@inf.ed.ac.uk

Andrew McCallum

University of Massachusetts
Amherst, MA 01003
USA
mccallum@cs.umass.edu

now
the essence of knowledge
Boston – Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is C. Sutton and A. McCallum, An Introduction to Conditional Random Fields, Foundation and Trends® in Machine Learning, vol 4, no 4, pp 267–373, 2011.

ISBN: 978-1-60198-572-9

Foundations and Trends® in Machine Learning

Volume 4 Issue 4, 2011

Editorial Board

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2011, Volume 4, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 4, No. 4 (2011) 267–373

DOI: 10.1561/2200000013

An Introduction to Conditional Random Fields

Charles Sutton[1] **and Andrew McCallum**[2]

[1] *School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK, csutton@inf.ed.ac.uk*
[2] *Department of Computer Science, University of Massachusetts, Amherst, MA, 01003, USA, mccallum@cs.umass.edu*

Abstract

Many tasks involve predicting a large number of variables that depend on each other as well as on other observed variables. Structured prediction methods are essentially a combination of classification and graphical modeling. They combine the ability of graphical models to compactly model multivariate data with the ability of classification methods to perform prediction using large sets of input features. This survey describes *conditional random fields*, a popular probabilistic method for structured prediction. CRFs have seen wide application in many areas, including natural language processing, computer vision, and bioinformatics. We describe methods for inference and parameter estimation for CRFs, including practical issues for implementing large-scale CRFs. We do not assume previous knowledge of graphical modeling, so this survey is intended to be useful to practitioners in a wide variety of fields.

Contents

1
Introduction

Fundamental to many applications is the ability to predict multiple variables that depend on each other. Such applications are as diverse as classifying regions of an image [49, 61, 69], estimating the score in a game of Go [130], segmenting genes in a strand of DNA [7], and syntactic parsing of natural-language text [144]. In such applications, we wish to predict an output vector $\mathbf{y} = \{y_0, y_1, \ldots, y_T\}$ of random variables given an observed feature vector $\mathbf{x}$. A relatively simple example from natural-language processing is part-of-speech tagging, in which each variable y_s is the part-of-speech tag of the word at position s, and the input $\mathbf{x}$ is divided into feature vectors $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Each $\mathbf{x}_s$ contains various information about the word at position s, such as its identity, orthographic features such as prefixes and suffixes, membership in domain-specific lexicons, and information in semantic databases such as WordNet.

One approach to this multivariate prediction problem, especially if our goal is to maximize the number of labels y_s that are correctly classified, is to learn an independent per-position classifier that maps $\mathbf{x} \mapsto y_s$ for each s. The difficulty, however, is that the output variables have complex dependencies. For example, in English adjectives do not usually follow nouns, and in computer vision, neighboring regions in an image tend to have similar labels. Another difficulty is that the output variables may represent a complex structure such as a parse tree, in which a choice of what grammar rule to use near the top of the tree can have a large effect on the rest of the tree.

A natural way to represent the manner in which output variables depend on each other is provided by graphical models. Graphical models — which include such diverse model families as Bayesian networks, neural networks, factor graphs, Markov random fields, Ising models, and others — represent a complex distribution over many variables as a product of local *factors* on smaller subsets of variables. It is then possible to describe how a given factorization of the probability density corresponds to a particular set of conditional independence relationships satisfied by the distribution. This correspondence makes modeling much more convenient because often our knowledge of the domain suggests reasonable conditional independence assumptions, which then determine our choice of factors.

Much work in learning with graphical models, especially in statistical natural-language processing, has focused on *generative* models that explicitly attempt to model a joint probability distribution $p(\mathbf{y}, \mathbf{x})$ over inputs and outputs. Although this approach has advantages, it also has important limitations. Not only can the dimensionality of $\mathbf{x}$ be very large, but the features may have complex dependencies, so constructing a probability distribution over them is difficult. Modeling the dependencies among inputs can lead to intractable models, but ignoring them can lead to reduced performance.

A solution to this problem is a *discriminative* approach, similar to that taken in classifiers such as logistic regression. Here we model the conditional distribution $p(\mathbf{y}|\mathbf{x})$ directly, which is all that is needed for classification. This is the approach taken by conditional random fields (CRFs). CRFs are essentially a way of combining the advantages of discriminative classification and graphical modeling, combining the ability to compactly model multivariate outputs $\mathbf{y}$ with the ability to leverage a large number of input features $\mathbf{x}$ for prediction. The advantage to a conditional model is that dependencies that involve only variables in $\mathbf{x}$ play no role in the conditional model, so that an accurate conditional model can have much simpler structure than a joint model. The difference between generative models and CRFs is thus exactly analogous to the difference between the naive Bayes and logistic regression classifiers. Indeed, the multinomial logistic regression model can be seen as the simplest kind of CRF, in which there is only one output variable.

There has been a large amount of interest in applying CRFs to many different problems. Successful applications have included text processing [105, 124, 125], bioinformatics [76, 123], and computer vision [49, 61]. Although early applications of CRFs used linear chains, recent applications of CRFs have also used more general graphical structures. General graphical structures are useful for predicting complex structures, such as graphs and trees, and for relaxing the independence assumption among entities, as in relational learning [142].

This survey describes modeling, inference, and parameter estimation using CRFs. We do not assume previous knowledge of graphical modeling, so this survey is intended to be useful to practitioners in a wide variety of fields. We begin by describing modeling issues in CRFs (Section 2), including linear-chain CRFs, CRFs with general graphical structure, and hidden CRFs that include latent variables. We describe how CRFs can be viewed both as a generalization of the well-known logistic regression procedure, and as a discriminative analogue of the hidden Markov model.

In the next two sections, we describe inference (Section 4) and learning (Section 5) in CRFs. In this context, *inference* refers both to the task of computing the marginal distributions of $p(\mathbf{y}|\mathbf{x})$ and to the related task of computing the maximum probability assignment $\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$. With respect to *learning*, we will focus on the parameter estimation task, in which $p(\mathbf{y}|\mathbf{x})$ is determined by parameters that we will choose in order to best fit a set of training examples $\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$. The inference and learning procedures are often closely coupled, because learning usually calls inference as a subroutine.

Finally, we discuss relationships between CRFs and other families of models, including other structured prediction methods, neural networks, and maximum entropy Markov models (Section 6).

1.1 Implementation Details

Throughout this survey, we strive to point out implementation details that are sometimes elided in the research literature. For example, we discuss issues relating to feature engineering (Section 2.5), avoiding numerical underflow during inference (Section 4.3), and the scalability of CRF training on some benchmark problems (Section 5.5).

Since this is the first of our sections on implementation details, it seems appropriate to mention some of the available implementations of CRFs. At the time of writing, a few popular implementations are:

CRF++	http://crfpp.sourceforge.net/
MALLET	http://mallet.cs.umass.edu/
GRMM	http://mallet.cs.umass.edu/grmm/
CRFSuite	http://www.chokkan.org/software/crfsuite/
FACTORIE	http://www.factorie.cc

Also, software for Markov Logic networks (such as Alchemy: http://alchemy.cs.washington.edu/) can be used to build CRF models. Alchemy, GRMM, and FACTORIE are the only toolkits of which we are aware that handle arbitrary graphical structure.

References

[1] S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325–343, 2000.

[2] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *International Conference on Machine Learning (ICML)*, 2003.

[3] G. Andrew and J. Gao, "Scalable training of l_1-regularized log-linear models," in *International Conference on Machine Learning (ICML)*, 2007.

[4] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, pp. 49–52, 1986.

[5] G. H. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, eds., *Predicting Structured Data*. MIT Press, 2007.

[6] T. Berg-Kirkpatrick, A. Bouchard-Côté, J. DeNero, and D. Klein, "Painless unsupervised learning with features," in *Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pp. 582–590.

[7] A. Bernal, K. Crammer, A. Hatzigeorgiou, and F. Pereira, "Global discriminative learning for higher-accuracy computational gene prediction," *PLoS Computational Biology*, vol. 3, no. 3, 2007.

[8] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2nd ed., 1999.

[9] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, vol. 24, no. 3, pp. 179–195, 1975.

[10] A. Blake, P. Kohli, and C. Rother, eds., *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 993, 2003.

[12] P. Blunsom and T. Cohn, "Discriminative word alignment with conditional random fields," in *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 65–72, 2006.

[13] L. Bottou, "Stochastic gradient descent examples on toy problems," 2010.

[14] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *International Conference on Computer Vision (ICCV)*, vol. 1, pp. 105–112, 2001.

[15] J. K. Bradley and C. Guestrin, "Learning tree conditional random fields," in *International Conference on Machine Learning (ICML)*, 2010.

[16] R. Bunescu and R. J. Mooney, "Collective information extraction with relational Markov networks," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[17] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-Newton matrices and their use in limited memory methods," *Mathematical Programming*, vol. 63, no. 2, pp. 129–156, 1994.

[18] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[19] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms using different performance metrics," Technical Report TR2005-1973, Cornell University, 2005.

[20] H. L. Chieu and H. T. Ng, "Named entity recognition with a maximum entropy approach," in *Conference on Natural Language Learning (CoNLL)*, pp. 160–163, 2003.

[21] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," in *Proceedings of the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, 2005.

[22] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun, "Map-reduce for machine learning on multicore," in *Advances in Neural Information Processing Systems 19*, pp. 281–288, MIT Press, 2007.

[23] S. Clark and J. R. Curran, "Parsing the WSJ using CCG and log-linear models," in *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, pp. 103–110, 2004.

[24] T. Cohn, "Efficient inference in large conditional random fields," in *European Conference on Machine Learning (ECML)*, pp. 606–613, Berlin, Germany, September 2006.

[25] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[26] P. J. Cowans and M. Szummer, "A graphical model for simultaneous partitioning and labeling," in *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[27] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, 2006.

[28] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *Journal of Machine Learning Research*, vol. 3, pp. 951–991, January 2003.

[29] A. Culotta, R. Bekkerman, and A. McCallum, "Extracting social networks and contact information from email and the web," in *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2004.

[30] A. Culotta and A. McCallum, "Confidence estimation for information extraction," in *Human Language Technology Conference (HLT)*, 2004.

[31] H. Daumé III, J. Langford, and D. Marcu, "Search-based structured prediction," *Machine Learning Journal*, 2009.

[32] H. Daumé III and D. Marcu, "Learning as search optimization: Approximate large margin methods for structured prediction," in *International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005.

[33] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *European Conference on Computer Vision (ECCV)*, 2010.

[34] J. V. Dillon and G. Lebanon, "Stochastic composite likelihood," *Journal of Machine Learning Research*, vol. 11, pp. 2597–2633, October 2010.

[35] G. Elidan, I. McGraw, and D. Koller, "Residual belief propagation: Informed scheduling for asynchronous message passing," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[37] J. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

[38] J. R. Finkel, A. Kleeman, and C. D. Manning, "Efficient, feature-based, conditional random field parsing," in *Annual Meeting of the Association for Computational Linguistics (ACL/HLT)*, pp. 959–967, 2008.

[39] K. Ganchev, J. Graca, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," Technical Report MS-CIS-09-16, University of Pennsylvania Department of Computer and Information Science, 2009.

[40] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, pp. 398–409, 1990.

[41] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[42] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Conference on Information and Knowledge Management (CIKM)*, 2005.

[43] A. Globerson, T. Koo, X. Carreras, and M. Collins, "Exponentiated gradient algorithms for log-linear structured prediction," in *International Conference on Machine Learning (ICML)*, 2007.

[44] J. Goodman, "Exponential priors for maximum entropy models," in *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2004.

[45] J. Graca, K. Ganchev, B. Taskar, and F. Pereira, "Posterior vs parameter sparsity in latent variable models," in *Advances in Neural Information Processing Systems 22*, (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 664–672, 2009.

[46] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[47] M. L. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 677–683, 2004.

[48] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *International Conference on Speech Communication and Technology*, 2005.

[49] X. He, R. S. Zemel, and M. A. Carreira-Perpiñián, "Multiscale conditional random fields for image labelling," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[50] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[51] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *BMC Bioinformatics*, vol. 6, no. Suppl 1, no. Suppl 1, 2005.

[52] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, and D. Schuurmans, "Semi-supervised conditional random fields for improved sequence segmentation and labeling," in *Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL)*, 2006.

[53] S. S. Keerthi and S. Sundararajan, "CRF versus SVM-struct for sequence labeling," Technical report, Yahoo! Research, 2007.

[54] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Annals of Mathematical Statistics*, vol. 23, pp. 462–466, 1952.

[55] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *International joint workshop on natural language processing in biomedicine and its applications*, pp. 70–75, Association for Computational Linguistics, 2004.

[56] P. Kohli, L. Ladickỳ, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision*, vol. 82, no. 3, no. 3, pp. 302–324, 2009.

[57] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[58] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, no. 2, pp. 498–519, 2001.

[59] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[60] A. Kulesza and F. Pereira, "Structured learning with approximate inference," in *Advances in Neural Information Processing Systems*, 2008.

[61] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," in *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[62] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, no. 2, pp. 179–201, 2006.

[63] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *International Conference on Machine Learning (ICML)*, 2001.

[64] J. Langford, A. Smola, and M. Zinkevich, "Slow learners are fast," in *Advances in Neural Information Processing Systems 22*, (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 2331–2339, 2009.

[65] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 504–513, 2010.

[66] Y. Le Cun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks, Tricks of the Trade*, Lecture Notes in Computer Science LNCS 1524, Springer Verlag, 1998.

[67] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[68] Y. LeCun, S. Chopra, R. Hadsell, R. Marc'Aurelio, and F.-J. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data*, (G. Bakir, T. Hofman, B. Schölkopf, A. Smola, and B. Taskar, eds.), MIT Press, 2007.

[69] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, 2001.

[70] W. Li and A. McCallum, "A note on semi-supervised learning using Markov random fields," 2004.

[71] P. Liang, H. Daumé III, and D. Klein, "Structure compilation: Trading structure for features," in *International Conference on Machine Learning (ICML)*, pp. 592–599, 2008.

[72] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *International Conference on Machine Learning (ICML)*, pp. 584–591, 2008.

[73] P. Liang, M. I. Jordan, and D. Klein, "Learning from measurements in exponential families," in *International Conference on Machine Learning (ICML)*, 2009.

[74] C.-J. Lin, R. C.-H. Weng, and S. Keerthi, "Trust region newton methods for large-scale logistic regression," in *Interational Conference on Machine Learning (ICML)*, 2007.

[75] B. G. Lindsay, "Composite likelihood methods," *Contemporary Mathematics*, pp. 221–239, 1988.

[76] Y. Liu, J. Carbonell, P. Weigele, and V. Gopalakrishnan, "Protein fold recognition using segmentation conditional random fields (SCRFs)," *Journal of Computational Biology*, vol. 13, no. 2, no. 2, pp. 394–406, 2006.

[77] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1150–1157, 1999.

[78] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "WinBUGS — a Bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, vol. 10, no. 4, no. 4, pp. 325–337, 2000.
[79] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
[80] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Conference on Natural Language Learning (CoNLL)*, (D. Roth and A. van den Bosch, eds.), pp. 49–55, 2002.
[81] G. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning of conditional random fields," in *Proceedings of Association of Computational Linguistics*, 2008.
[82] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, no. 2, pp. 313–330, 1993.
[83] A. McCallum, "Efficiently inducing features of conditional random fields," in *Conference on Uncertainty in AI (UAI)*, 2003.
[84] A. McCallum, K. Bellare, and F. Pereira, "A conditional random field for discriminatively-trained finite-state string edit distance," in *Conference on Uncertainty in AI (UAI)*, 2005.
[85] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *International Conference on Machine Learning (ICML)*, pp. 591–598, San Francisco, CA, 2000.
[86] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Seventh Conference on Natural Language Learning (CoNLL)*, 2003.
[87] A. McCallum, K. Schultz, and S. Singh, "FACTORIE: Probabilistic programming via imperatively defined factor graphs," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.
[88] A. McCallum and B. Wellner, "Conditional models of identity uncertainty with application to noun coreference," in *Advances in Neural Information Processing Systems 17*, (L. K. Saul, Y. Weiss, and L. Bottou, eds.), pp. 905–912, Cambridge, MA: MIT Press, 2005.
[89] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's "belief propagation" algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, no. 2, pp. 140–152, 1998.
[90] S. Miller, J. Guinness, and A. Zamanian, "Name tagging with word clusters and discriminative training," in *HLT-NAACL 2004: Main Proceedings*, (D. Marcu, S. Dumais, and S. Roukos, eds.), pp. 337–342, Boston, Massachusetts, USA: Association for Computational Linguistics, May 2–May 7 2004.
[91] T. P. Minka, "The EP energy function and minimization schemes," Technical report, 2001.
[92] T. P. Minka, "A comparsion of numerical optimizers for logistic regression," Technical report, 2003.
[93] T. P. Minka, "Discriminative models, not discriminative training," Technical Report MSR-TR-2005-144, Microsoft Research, October 2005.

[94] T. P. Minka, "Divergence measures and message passing," Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

[95] I. Murray, "Advances in Markov chain Monte Carlo methods," PhD thesis, Gatsby computational neuroscience unit, University College London, 2007.

[96] I. Murray, Z. Ghahramani, and D. J. C. MacKay, "MCMC for doubly-intractable distributions," in *Uncertainty in Artificial Intelligence (UAI)*, pp. 359–366, AUAI Press, 2006.

[97] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *International Conference on Machine Learning (ICML)*, 2004.

[98] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems 14*, (T. G. Dietterich, S. Becker, and Z. Ghahramani, eds.), pp. 841–848, Cambridge, MA: MIT Press, 2002.

[99] N. Nguyen and Y. Guo, "Comparisons of sequence labeling algorithms and extensions," in *International Conference on Machine Learning (ICML)*, 2007.

[100] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[101] S. Nowozin and C. H. Lampert, "Structured prediction and learning in computer vision," *Foundations and Trends in Computer Graphics and Vision*, vol. 6, no. 3-4, no. 3-4, 2011.

[102] C. Pal, C. Sutton, and A. McCallum, "Sparse forward-backward using minimum divergence beams for fast training of conditional random fields," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[103] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[104] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in *International Conference on Computational Linguistics (COLING)*, pp. 562–568, 2004.

[105] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," in *Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2004.

[106] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[107] Y. Qi, M. Szummer, and T. P. Minka, "Bayesian conditional random fields," in *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[108] Y. Qi, M. Szummer, and T. P. Minka, "Diagram structure recognition by Bayesian conditional random fields," in *International Conference on Computer Vision and Pattern Recognition*, 2005.

[109] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1104, 2005.

[110] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden-state conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[111] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, no. 2, pp. 257–286, 1989.

[112] N. Ratliff, J. A. Bagnell, and M. Zinkevich, "Maximum margin planning," in *International Conference on Machine Learning*, July 2006.

[113] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1–2, no. 1–2, pp. 107–136, 2006.

[114] S. Riezler, T. King, R. Kaplan, R. Crouch, J. T. Maxwell III, and M. Johnson, "Parsing the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[115] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[116] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[117] D. Rosenberg, D. Klein, and B. Taskar, "Mixture-of-parents maximum entropy Markov models," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[118] D. Roth and W. Yih, "Integer linear programming inference for conditional random fields," in *International Conference on Machine Learning (ICML)*, pp. 737–744, 2005.

[119] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 23, no. 3, no. 3, pp. 309–314, 2004.

[120] E. F. T. K. Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: Chunking," in *Proceedings of CoNLL-2000 and LLL-2000*, 2000. See http://lcg-www.uia.ac.be/~erikt/research/np-chunking.html.

[121] E. F. T. K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of CoNLL-2003*, (W. Daelemans and M. Osborne, eds.), pp. 142–147, Edmonton, Canada, 2003.

[122] S. Sarawagi and W. W. Cohen, "Semi-Markov conditional random fields for information extraction," in *Advances in Neural Information Processing Systems 17*, (L. K. Saul, Y. Weiss, and L. Bottou, eds.), pp. 1185–1192, Cambridge, MA: MIT Press, 2005.

[123] K. Sato and Y. Sakakibara, "RNA secondary structural alignment with conditional random fields," *Bioinformatics*, vol. 21, pp. ii237–242, 2005.

[124] B. Settles, "Abner: An open source tool for automatically tagging genes, proteins, and other entity names in text," *Bioinformatics*, vol. 21, no. 14, no. 14, pp. 3191–3192, 2005.

[125] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Conference on Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL)*, pp. 213–220, 2003.

[126] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM," in *International Conference on Machine Learning (ICML)*, 2007.

[127] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation," in *European Conference on Computer Vision (ECCV)*, 2006.

[128] P. Singla and P. Domingos, "Discriminative training of Markov logic networks," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 868–873, Pittsburgh, PA, 2005.

[129] F. K. Soong and E.-F. Huang, "A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1991.

[130] D. H. Stern, T. Graepel, and D. J. C. MacKay, "Modelling uncertainty in the game of go," in *Advances in Neural Information Processing Systems 17*, (L. K. Saul, Y. Weiss, and L. Bottou, eds.), pp. 1353–1360, Cambridge, MA: MIT Press, 2005.

[131] I. Sutskever and T. Tieleman, "On the convergence properties of contrastive divergence," in *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[132] C. Sutton, "Efficient Training Methods for Conditional Random Fields," PhD thesis, University of Massachusetts, 2008.

[133] C. Sutton and A. McCallum, "Collective segmentation and labeling of distant entities in information extraction," in *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004.

[134] C. Sutton and A. McCallum, "Piecewise training of undirected models," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[135] C. Sutton and A. McCallum, "Improved dynamic schedules for belief propagation," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[136] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, (L. Getoor and B. Taskar, eds.), MIT Press, 2007.

[137] C. Sutton and A. McCallum, "Piecewise training for structured prediction," *Machine Learning*, vol. 77, no. 2–3, no. 2–3, pp. 165–194, 2009.

[138] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, pp. 693–723, March 2007.

[139] C. Sutton and T. Minka, "Local training and belief propagation," Technical Report TR-2006-121, Microsoft Research, 2006.

[140] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *International Conference on Machine Learning (ICML)*, 2004.

[141] C. Sutton, M. Sindelar, and A. McCallum, "Reducing weight undertraining in structured discriminative learning," in *Conference on Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL)*, 2006.

[142] B. Taskar, P. Abbeel, and D. Koller, "Discriminative probabilistic models for relational data," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

[143] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Advances in Neural Information Processing Systems 16*, (S. Thrun, L. Saul, and B. Schölkopf, eds.), Cambridge, MA: MIT Press, 2004.

[144] B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning, "Max-margin parsing," in *Empirical Methods in Natural Language Processing (EMNLP04)*, 2004.

[145] B. Taskar, S. Lacoste-Julien, and D. Klein, "A discriminative matching approach to word alignment," in *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pp. 73–80, 2005.

[146] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *HLT-NAACL*, 2003.

[147] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Interational Conference on Machine Learning (ICML)*, ICML '04, 2004.

[148] P. Viola and M. Narasimhan, "Learning to extract information from semi-structured text using a discriminative context free grammar," in *Proceedings of the ACM SIGIR*, 2005.

[149] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. Murphy, "Accelerated training of conditional random fields with stochastic meta-descent," in *International Conference on Machine Learning (ICML)*, pp. 969–976, 2006.

[150] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, no. 1-2, pp. 1–305, 2008.

[151] M. J. Wainwright, "Estimating the wrong Markov random field: Benefits in the computation-limited setting," in *Advances in Neural Information Processing Systems 18*, (Y. Weiss, B. Schölkopf, and J. Platt, eds.), Cambridge, MA: MIT Press, 2006.

[152] M. J. Wainwright, T. Jaakkola, and A. S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Transactions on Information Theory*, vol. 45, no. 9, no. 9, pp. 1120–1146, 2003.

[153] H. Wallach, "Efficient training of conditional random fields," M.Sc. thesis, University of Edinburgh, 2002.

[154] M. Welling and S. Parise, "Bayesian random fields: The Bethe-Laplace approximation," in *Uncertainty in Artificial Intelligence (UAI)*, 2006.

[155] M. Wick, K. Rohanimanesh, A. Culotta, and A. McCallum, "SampleRank: Learning preferences from atomic gradients," in *Neural Information Processing Systems (NIPS) Workshop on Advances in Ranking*, 2009.

[156] M. Wick, K. Rohanimanesh, A. McCallum, and A. Doan, "A discriminative approach to ontology alignment," in *International Workshop on New Trends in Information Integration (NTII)*, 2008.

[157] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," Technical Report TR2004-040, Mitsubishi Electric Research Laboratories, 2004.

[158] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, July 2005.

[159] C.-N. Yu and T. Joachims, "Learning structural svms with latent variables," in *International Conference on Machine Learning (ICML)*, 2009.

[160] J. Yu, S. V. N. Vishwanathan, S. Güunter, and N. N. Schraudolph, "A quasi-Newton approach to nonsmooth convex optimization problems in machine learning," *Journal of Machine Learning Research*, vol. 11, pp. 1145–1200, March 2010.

[161] Y. Zhang and C. Sutton, "Quasi-Newton Markov chain Monte Carlo," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.