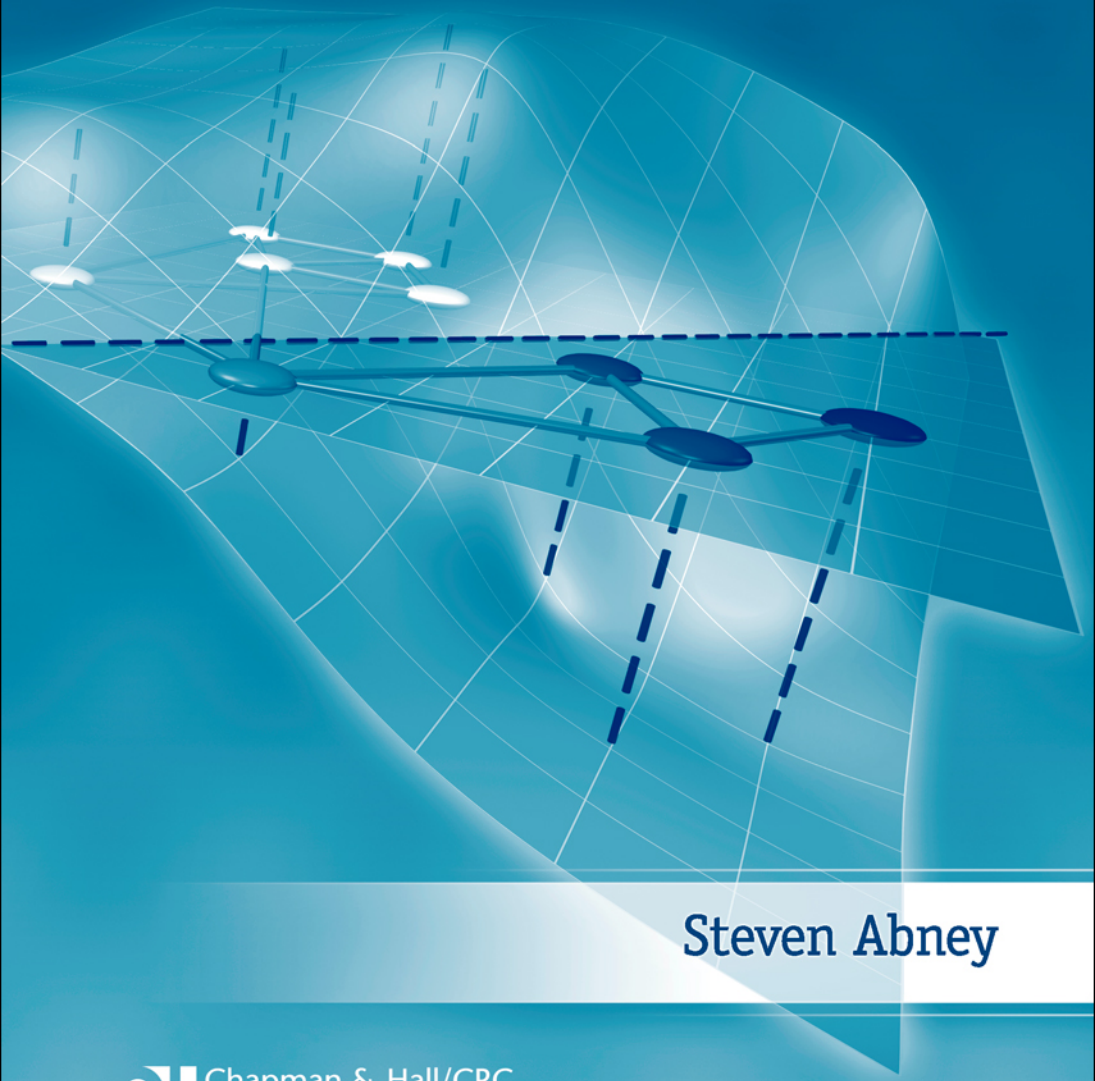
Computer Science and Data Analysis Series

Semisupervised Learning for Computational Linguistics



Steven Abney

Chapman & Hall/CRC
Taylor & Francis Group

Computer Science and Data Analysis Series

Semisupervised Learning for Computational Linguistics

Chapman & Hall/CRC

Computer Science and Data Analysis Series

The interface between the computer and statistical sciences is increasing, as each discipline seeks to harness the power and resources of the other. This series aims to foster the integration between the computer sciences and statistical, numerical, and probabilistic methods by publishing a broad range of reference works, textbooks, and handbooks.

Published Titles

Bayesian Artificial Intelligence
Kevin B. Korb and Ann E. Nicholson

Pattern Recognition Algorithms for Data Mining
Sankar K. Pal and Pabitra Mitra

Exploratory Data Analysis with MATLAB®
Wendy L. Martinez and Angel R. Martinez

Clustering for Data Mining: A Data Recovery Approach
Boris Mirkin

Correspondence Analysis and Data Coding with Java and R
Fionn Murtagh

R Graphics
Paul Murrell

Design and Modeling for Computer Experiments
Kai-Tai Fang, Runze Li, and Agus Sudjianto

Semisupervised Learning for Computational Linguistics
Steven Abney

Computer Science and Data Analysis Series

Semisupervised Learning for Computational Linguistics

Steven Abney
University of Michigan
Ann Arbor, U.S.A.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an **informa** business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-559-7 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Abney, Steven P.
Semisupervised learning in computational linguistics / Steven Abney.
p. cm. -- (Computer science and data analysis series)
ISBN 978-1-58488-559-7 (alk. paper)
1. Computational linguistics--Study and teaching (Higher) I. Title. II. Series.

P98.3.A26 2007
410.285--dc22 2007022858

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Steven Abney

Semisupervised Learning for Computational Linguistics

CRC PRESS
Boca Raton Ann Arbor London Tokyo

Preamble

The primary audience for this book is students, researchers, and developers in computational linguistics who are interested in applying or advancing our understanding of semisupervised learning methods for natural language processing. The problem of semisupervised learning arose almost immediately when computational linguists began exploring statistical and machine learning methods seriously in the late 1980s and early 1990s. In fact, language applications – particularly text classification and information extraction – have provided a major impetus for interest in semisupervised learning in the machine learning community.

Statistical methods that combine labeled and unlabeled data go back at least to the 1960s, and theoretical understanding has advanced quickly over the last few years; but the rate of advancements has made it difficult for non-specialists to keep abreast of them. Those computational linguists whose interest in semisupervised learning is more practical and empirical would benefit from an accessible presentation of the state of the theory. For students, the need for an accessible presentation is urgent.

The purpose of the book is to provide students and researchers a broad and accessible presentation of what is currently known about semisupervised learning, including both the theory and linguistic applications. The background assumed is what can be reasonably expected of any graduate student (or even an advanced undergraduate) who has taken introductory courses in natural language processing that include statistical methods – concretely, the material contained in Jurafsky & Martin [119] and Manning & Schütze [141].

It is desirable that the book be self-contained. Consequently, its coverage will overlap somewhat with standard texts in machine learning. This is unavoidable, given that the target audience is not assumed to have background in machine learning beyond what is contained in the texts just mentioned, and given that semisupervised learning cannot be seriously tackled without understanding the methods for supervised and unsupervised learning that it builds on. My approach has been to treat only those topics in supervised and unsupervised learning that are necessary for understanding semisupervised methods, and to aim for intuitive understanding rather than rigor and completeness – again, except insomuch as a rigorous treatment is required for understanding the main topic, the semisupervised case. In short, the book does cover a number of topics that are found in general introductions to machine learning; but if viewed as a general introduction, it will seem eclectic in coverage and intuitive in treatment. I do not see this necessarily as a flaw.

I find that my own interests often run beyond the areas where I have solid foundations, and an intuitive overview gives me motivation to go back and fill in those foundations. I hope that students of computational linguistics who come with an interest in semisupervised problems but without a general training in machine learning will, above all, find the main topic accessible, but will also acquire a framework and motivation for more systematic study of machine learning.

Although the book is written with computational linguists in mind, I hope that it will also be of interest to students of machine learning. Simple text classification tasks are now familiar in the machine learning literature, but fewer machine learning researchers are aware of the variety of other linguistic applications. Moreover, linguistic applications have characteristic properties that differ in interesting ways from applications that have been the traditional focus in machine learning, and can suggest new questions for theoretical study. For example, natural language problems often have attributes with large sets of discrete values with highly skewed distributions (that is, word-valued attributes), or large sparse spaces of real-valued attributes (numeric attributes indexed by words), or learning targets that are neither discrete classes nor real values, but rather structures (text spans or parse trees).

Perhaps a few readers from even further afield will find the book useful. I have benefited from a book on clustering written for chemists [145], and I would be pleased if researchers from areas well outside of natural language processing find something useful here. Semisupervised learning is a topic of broad applicability. It has already been applied to image processing [98], bioinformatics, and security assessment [177], to name a few examples. Further applications are limited only by imagination.

Acknowledgments

First and foremost, I would like to thank Mark Abney for producing the cover and most of the illustrations. He has done a terrific job; without him the book would have taken much longer and been a good deal less attractive.

I would also like to thank my students and colleagues for reading an earlier draft and giving me indispensable feedback, especially Güneş Erkan, Jessica Hullman, Kevin McGowan, Terrence Szymanski, Richmond Thomason, Li Yang, and an anonymous reviewer.

Finally, I would like to thank my family, Marie Carmen, Anneliese, and Nina, for their patience and support while this book has been in the making.

Contents

1

Introduction

1.1 A brief history

1.1.1 Probabilistic methods in computational linguistics

Computational linguistics seeks to describe methods for natural language processing, that is, for processing human languages by automatic means. Since the advent of electronic computers in the late 1940s, human language processing has been an area of active research; machine translation in particular attracted early interest. Indeed, the inspiration for computing machines was the creation of a thinking automaton, a *machina sapiens*, and language is perhaps the most distinctively human cognitive capacity.

In early work on artificial intelligence, there was something of a competition between discrete, "symbolic" reasoning and stochastic systems, particularly neural nets. But the indispensability of a firm probabilistic basis for dealing with uncertainty was soon recognized. In computational linguistics, by contrast, the presumption of the sufficiency of grammatical and logical constraints, supplemented perhaps by ad hoc heuristics, was much more tenacious.

When the field recognized the need for probabilistic methods, the shift was sudden and dramatic. It is probably fair to identify the birth of awareness with the appearance in 1988 of two papers on statistical part-of-speech tagging, one by Church [44] and one by DeRose [75]. These were not the first papers that proposed stochastic methods for part of speech disambiguation, but they were the first in prominent venues in computational linguistics, and it is no exaggeration to say that the field was reshaped within a decade.

The main barrier to progress in natural language processing at the time was the brittleness of manually constructed systems. The dominant issues were encapsulated under the rubrics of ambiguity resolution, portability, and robustness. The primary method for **ambiguity resolution** was the use of semantic constraints, but they were often either too loose, leaving a large number of viable analyses, or else too strict, ruling out the correct analysis. Well-founded and automatic means for softening constraints and resolving ambiguities were needed. **Portability** meant in particular automatic means for adapting to variability across application domains. **Robustness** covers both the fact that input to natural language systems is frequently errorful, and

also the fact that, in Sapir's terms, "all grammars leak" [201]. No manually constructed description of language is complete.

Together, these issues point to the need for automatic learning methods, and explain why the penetration of probabilistic methods, and machine learning in particular, was so rapid. Computational linguistics has now become inseparable from machine learning.

1.1.2 Supervised and unsupervised training

The probabilistic models used by Church and DeRose in the papers just cited were Hidden Markov Models (HMMs), imported from the speech recognition community. An HMM describes a probabilistic process or automaton that generates sequences of states and parallel sequences of output symbols. Commonly, a sequence of output symbols represents a sentence of English or of some other natural language. An HMM, or any model, that defines probabilities of word sequences (that is, sentences) of a natural language is known as a **language model**.

The probabilistic automaton defined by an HMM may be in some number of distinct **states**. The automaton begins by choosing a state at random. Then it chooses a symbol to emit, the choice being sensitive to the state. Next it chooses a new state, emits a symbol from that state, and the process repeats. Each choice is stochastic – that is, probabilistic. At each step, the automaton makes its choice at random from a distribution over output symbols or next states, as the case may be. Which distribution it uses at any point is completely determined by the kind of choice, either **emission** of an output symbol or **transition** to a new state, and the identity of the current state. The actual **model** consists in a collection of numeric values, one for each possible transition or emission, representing the probability that the automaton chooses that particular transition or emission when making one of its stochastic choices.

Learning an HMM is straightforward if one is provided with **labeled data**, meaning state sequences paired with output sequences. Each sequence pair is a record of the stochastic choices made by the automaton. To estimate the probability that the automaton will choose a particular value x when faced with a stochastic choice of type T, one can simply count how often the automaton actually chose x when making a choice of type T in the record of previous computations, that is, in the labeled data. If sufficient labeled data is available, the model can be estimated accurately in this way.

Church and DeRose applied HMMs to the problem of part-of-speech tagging by identifying the states of the automaton with parts of speech. The automaton generates a sequence of parts of speech, and emits a word for each part of speech. The result is a **tagged text**, which is a text in which each word is annotated with its part of speech.

Supervised learning of an HMM for part-of-speech tagging is quite effective; HMM taggers for English generally have an error rate of 3.5 to 4 percent.

Their effectiveness was what brought probabilistic models to the attention of computational linguists, as already mentioned.

1.1.3 Semisupervised learning

Creating sufficient labeled data can be very time-consuming. Obtaining the output sequences is not difficult: English texts are available in great quantity. What is time-consuming is creating the state sequences. One must essentially annotate each word of the texts with the state of the HMM from which it was emitted. For this reason, one would like to have a method for learning a model from **unlabeled data**, which, in this case, consists of simple English texts without state annotations. A learning method that uses unlabeled data is known as an **unsupervised** learning method, in contrast to **supervised** learning methods, which use labeled data.

An unsupervised learning method for HMMs has long been known, called the forward-backward algorithm. It is a special case of the Expectation-Maximization (EM) algorithm. It is widely used and effective in speech recognition. For example, it is used to estimate acoustic models from unlabeled data. For an acoustic model, unlabeled data consists of text paired with the speech signal resulting from reading the text, but without any annotation regarding the sequence of states that the model passes through while generating the speech from the text.

However, unsupervised learning turned out not to be very effective for part-of-speech tagging. The forward-backward algorithm is an iterative algorithm that begins with some initial model and improves it by repeated passes through the unlabeled data. The question of the effectiveness of unsupervised learning was posed by Elworthy [83] in the following form. If one uses labeled data to obtain the initial model for forward-backward training on unlabeled data, how many iterations of forward-backward training should one do for a given amount of labeled seed data? The answer was essentially zero iterations if one had more than a tiny amount of labeled data. (Merialdo [153] came to similar conclusions.) Intuitively, the states that are learned by an automaton trained on unlabeled data do not correspond at all well to linguistically motivated parts of speech.

The solution that emerged in the case of part-of-speech tagging was to use additional constraints. Specifically, if one uses unlabeled data and forward-backward training, but restricts words to take on only the parts of speech that are allowed for them by a dictionary, the results are comparable to using labeled data [62].

But despite its ineffectiveness for part-of-speech tagging, the idea of learning from a mixture of labeled and unlabeled data remained potent, and it has come to constitute the canonical setting for **semisupervised learning**. (One can view the dictionary constraints that proved more effective for tagging as providing partially labeled data, hence a variant of semisupervised learning. This will be discussed in more detail later.)

Subsequent work in computational linguistics led to development of alternative algorithms for semisupervised learning, the algorithm of Yarowsky [239] being a prominent example. These algorithms were developed specifically for the sorts of problems that arise frequently in computational linguistics: problems in which there is a linguistically correct answer, and large amounts of unlabeled data, but very little labeled data. Unlike in the example of acoustic modeling, classic unsupervised learning is inappropriate, because not just any way of assigning classes will do. The learning method is largely unsupervised, because most of the data is unlabeled, but the labeled data is indispensable, because it provides the only characterization of the linguistically correct classes.

The algorithms just mentioned turn out to be very similar to an older learning method known as **self-training** that was unknown in computational linguistics at the time. For this reason, it is more accurate to say that they were rediscovered, rather than invented, by computational linguists. Until very recently, most prior work on semisupervised learning has been little known even among researchers in the area of machine learning. One goal of the present volume is to make the prior and also the more recent work on semisupervised learning more accessible to computational linguists.

Shortly after the rediscovery of self-training in computational linguistics, a method called **co-training** was invented by Blum and Mitchell [21], machine-learning researchers working on text classification. Self-training and co-training have become popular and widely employed in computational linguistics; together they account for all but a fraction of the work on semisupervised learning in the field. We will discuss them in the next chapter. In the remainder of this chapter, we give a broader perspective on semisupervised learning, and lay out the plan of the rest of the book.

1.2 Semisupervised learning

1.2.1 Major varieties of learning problem

There are five types of learning problem that have received the preponderance of attention in machine learning. The first four are all cases of **function estimation**, grouped along two dimensions: whether the learning task is supervised or unsupervised, and whether the variable to be predicted is nominal or real-valued.

Classification involves supervised learning of a function $f(x)$ whose value is nominal, that is, drawn from a finite set of possible values. The learned function is called a classifier. It is given **instances** x of one or another class, and it must determine which class each instance belongs to; the value $f(x)$ is the classifier's prediction regarding the class of the instance. For example, an

instance might be a particular word in context, and the classification task is to determine its part of speech. The learner is given labeled data consisting of a collection of instances along with the correct answer, that is, the correct class label, for each instance.

The unsupervised counterpart to classification is **clustering**. The goal in clustering is also to assign instances to classes, but the clustering algorithm is given only the instances, not the correct answers for any of them. (In clustering, the instances are usually called **data points** and the classes are called **clusters**.) The primary difference between classification and clustering is not the task to be performed, but the sort of data that is given to the learner as input; in particular, whether the data is labeled or not.

The remaining two function estimation tasks involve estimation of a function that takes on real values, instead of values from a finite range. The supervised version is called **regression**; it differs from classification only in that the function to be learned takes on real values. Unsupervised learning of a real-valued function can be viewed as **density estimation**. The learner is given an unlabeled set of training data, consisting of a finite sample of data points from a multi-dimensional space, and the goal is to learn a function $f(x)$ assigning a real value to every point in the space; the function is interpreted as (proportional to) a probability density.

Finally, we mentioned a fifth setting that does not fall under function estimation. This fifth setting is **reinforcement learning**. In reinforcement learning, the learner receives a stream of data from sensors, and its "answers" consist in actions, in the form of commands sent to actuators. There is, additionally, a reward signal that is to be maximized (over the long run). There are at least two significant ways this differs from the four function estimation settings. First is the sequential nature of the inputs. Even if we assume discrete time, there are temporal dependencies that cannot be ignored: in particular, actions have time-delayed effects on sensors and reward. Second is the indirect nature of the supervision. The reward signal provides information about the relative value of different actions, but it is much less direct than simply providing the correct answer, as in classification.

Semisupervised learning generalizes supervised and unsupervised learning. The generalization is easiest to see with classification and clustering. As already mentioned, classification and clustering involve essentially the same task and the same inputs; they differ primarily in whether the training data is labeled or not. (They also differ in the way they are evaluated, but the difference in evaluation is a consequence of the difference in the kind of training data – more on that later.) The obvious generalization is to give the learner labels for *some* of the training data. At one extreme, all of the data is labeled, and the task is classification, and at the other extreme, none of the data is labeled, and the task is clustering. The mixed labeled/unlabeled setting is indeed the canonical case for semisupervised learning, and it will be our main interest.

At the same time, a mix of labeled and unlabeled information is only one

way of providing a learner with partial information about the labels for training data. Many semisupervised learning methods work with alternate kinds of partial information, such as a handful of reliable rules for labeling instances, or constraints limiting the candidate labels for particular instances. We will also consider these extensions of the canonical setting. In principle, the kind of indirect information about labels found in reinforcement learning qualify it as a kind of semisupervised learning, but the indirect-information aspect of reinforcement learning is difficult to disentangle from the temporal dependencies, and the connection between reinforcement learning and other semisupervised approaches remains obscure; it lies beyond the scope of the present work.

1.2.2 Motivation

For most learning tasks of interest, it is easy to obtain samples of unlabeled data. For many language learning tasks, for example, the World Wide Web can be seen as a large collection of unlabeled data. By contrast, in most cases, the only practical way to obtain labeled data is to have subject-matter experts manually annotate the data, an expensive and time-consuming process.

The great advantage of unsupervised learning, such as clustering, is that it requires no labeled training data. The disadvantage has already been mentioned: under the best of circumstances, one might hope that the learner would recover the correct clusters, but hardly that it could correctly label the clusters. In many cases, even the correct clusters are too much to hope for. To say it another way, unsupervised learning methods rarely perform well if evaluated by the same yardstick used for supervised learners. If we expect a clustering algorithm to predict the labels in a labeled test set, without the advantage of labeled training data, we are sure to be disappointed.

The advantage of supervised learning algorithms is that they do well at the harder task: predicting the true labels for test data. The disadvantage is that they only do well if they are given enough labeled training data, but producing sufficient quantities of labeled data can be very expensive in manual effort.

The aim of semisupervised learning is to have our cake and eat it, too. Semisupervised learners take as input unlabeled data and a limited source of label information, and, if successful, achieve performance comparable to that of supervised learners at significantly reduced cost in manual production of training data.

We intentionally used the vague phrase "a limited source of label information." One source of label information is obviously labeled data, but there are alternatives. We will consider at least the following sources of label information:

- labeled data
- a seed classifier

- limiting the possible labels for instances without determining a unique label
- constraining pairs of instances to have the same, but unknown, label (co-training)
- intrinsic label definitions
- a budget for labeling instances selected by the learner (active learning)

One of the grand aims of computational linguistics is unsupervised learning of natural language. From a psychological perspective, it is widely accepted that explicit instruction plays little part in human language learning, and from a technological perspective, a completely autonomous system is more useful than one that requires manual guidance. Yet, in contradiction to the characterization sometimes given of the goal of unsupervised learning, the goal of unsupervised language learning is not the recovery of arbitrary "interesting" structure, but rather the acquisition of the correct target language. On the face of it, learning a target classification – much less an entire natural language – without labeled data hardly seems possible.

Semisupervised learning may provide the beginning of an account. If a kernel of labeled data can be acquired through unsupervised learning, semisupervised learning might be used to extend it to a complete solution. Something along these lines appears to characterize human language acquisition: in the psycholinguistic literature, *bootstrapping* refers to the process by which an initial kernel of language is acquired by explicit instruction, in the form, for example, of naming an object while drawing a child's attention to it. The processes by which that kernel is extended to the entirety of the language are thought to be different; distributional regularities of linguistic forms, rather than direct connections to the physical world, seem to play a large role. Semisupervised learning methods provide possible characterizations of the process of extending the initial kernel.

1.2.3 Evaluation

With regard to evaluation, semisupervised algorithms are like supervised algorithms. The basic measure of success is classification performance on an unseen test set, used as an estimate of generalization error.

But in addition to measuring absolute performance, one would also like to measure the benefit obtained by the addition of unlabeled data. The most general way to pose the question is the level of performance as a function of human effort. More concretely, one considers prediction rule quality as a function of the number of labeled instances and the number of unlabeled instances. Two questions are of particular interest: (1) for a fixed number of labeled instances (i.e., a fixed annotation budget), how much improvement is obtainable as the number of unlabeled instances grows without bound; and

(2) for a fixed target level of performance, what is the minimum number of labeled instances needed to achieve it, as the number of unlabeled instances grows without bound.

1.2.4 Active learning

One way of characterizing the overarching goal of semisupervised learning is to use unlabeled data to amplify the information gained from a manually created seed. We focus almost exclusively on "batch" learning, in which the seed and a population of unlabeled data are given in advance. A natural next question is whether a better effort-performance curve can be obtained in an interactive setting, for example, by selecting instances to be labeled, and interleaving learning with labeling. Interactive semisupervised learning is called **active learning**. It lies beyond the scope of the current work.

1.3 Organization and assumptions

1.3.1 Leading ideas

Semisupervised learning methods have sprung up independently in several different areas, usually as modifications of existing algorithms. For example, if one's interest is classification, it is natural to ask how to modify a classifier-learning algorithm to make use of unlabeled data. Conversely, if one's interest is clustering, it is natural to ask how to make use of manually labeled examples, either to assign names to otherwise anonymous clusters, or to constrain the algorithm to produce clusters that are consistent with the manual labels.

We organize semisupervised algorithms by the leading idea that each is based on. These are the leading ideas, in our view:

- **Self-training** (chapters 2–3 and 8). If one comes from the perspective of supervised learning, and asks how unlabeled instances might be put to use in addition to labeled instances, a natural idea is to train a classifier on the labeled instances, apply it to the unlabeled instances, and take its predictions at face value, at least in those cases where its predictions are most confident. A new classifier is trained on the extended set of labeled instances, and the process repeats. This approach is known as **self-training**.

- **Cluster and label** (chapter 7). Coming to semisupervised learning from the perspective of unsupervised learning, a natural idea is to apply a clustering algorithm to the unlabeled data (one can also strip the labels from the labeled data and throw it in as well), and then use the labeled data to "name" the clusters. A cluster is associated with whichever

label occurs most frequently on labeled instances in the cluster, and the prediction for an unlabeled instance is determined by the cluster that it is assigned to.

- **Application of "missing values" techniques** (chapter 8). The problem of missing values in a data set is very familiar in statistics. It is natural to think of unlabeled data as data with missing values for the dependent variable (that is, the class label), and apply a method for filling in missing information using a **generative model**. The canonical example is the Expectation-Maximization (EM) algorithm. The earliest literature on semisupervised learning falls under this rubric.

- **Label propagation in graphs** (chapter 10). Clustering algorithms are typically based on a similarity metric; clusters are defined to be groups of similar instances. A postulate sometimes called the **cluster hypothesis** is that similar instances have similar labels, or in geometric terms, that proximate instances have similar labels. A similarity function can be represented as a weighted graph, in which instances are nodes and edges are weighted by the similarity of the instances they connect. Two nodes connected by a heavily weighted edge should have the same label. An algorithmic correlate is to propagate labels along heavily weighted edges. Geometrically, one can view the graph as a fabric whose interior consists of unlabeled instances and whose boundary consists of labeled instances. The elevation of the fabric at a given point represents the label of that point, and the effect of propagation is to interpolate from the fixed boundary across the interior of the graph.

- **Boundaries in low density regions** (chapters 5–6). The contrapositive of the cluster hypothesis is what we might call the **separation hypothesis**: the idea that different labels imply distant data points, which is to say, that boundaries between classes lie where the data is sparse. **Transductive** maximum-margin methods, such as transductive Support Vector Machines (SVMs) and transductive boosting, can be understood in those terms. The goal is to find a linear inter-class boundary with a large *margin*, which is to say, a large distance to the nearest data points. Instead of looking for natural clusters, one looks for natural boundaries, but otherwise the approach is very similar to cluster-and-label. Natural boundaries are found without regard to labels, and the labeled instances are used to determine labels for the resulting regions.

- **Constraint- and agreement-driven learning** (chapters 2–3 and 9). In a sense, all semisupervised learning is driven by constraints. Sufficiently restrictive constraints can be almost as good as labels for unlabeled data – and in some cases even better, inasmuch as a constraint applies to the entire population of instances, whereas a label on an instance applies only to that instance. We have mentioned how graph methods

translate similarities into soft constraints. A particularly salient class of constraints is agreement constraints. For example, in **co-training**, the learner is given two independent "views" of the data, and constructs one classifier for each view, under the constraint that the classifiers agree in their predictions on the unlabeled data. Effectively, instances come in pairs that are constrained to have the same label. The learner does not know what the label is, but does know that it is the same for both members of the pair. A non-algorithmic way of enforcing agreement is via a **random field** that penalizes disagreement.

- **Spectral methods** (chapters 11–12). One can build on the idea of interpolation that emerges from label propagation by using a "standing wave" to interpolate across the graph. Pursuing this idea leads to deep connections among apparently disparate ideas, including the cluster hypothesis and label propagation, "mincut" boundary-oriented methods, and random fields.

The plan of the book more or less follows the list of leading ideas just given, with a couple of rearrangements for the sake of a smoother line of development. We begin, in chapters 2 and 3, with a discussion of the semisupervised methods that are already well known in computational linguistics, namely, self-training and co-training. We turn then to methods that come from the machine learning literature, beginning in chapter 4 with an introduction to classification, including some detail on the topic of decision boundaries. The discussion of boundary-oriented methods follows naturally at that point (chapters 5–6). Then we turn to clustering in chapter 7, followed by discussion of the EM algorithm and related generative methods in chapter 8. There are connections between co-training and the generative methods of chapter 8, so the chapter on agreement methods is placed next (chapter 9). Finally, the strand of graph-based methods, begun in the chapter on clustering, is picked up in chapter 10, which concerns label propagation, and in chapters 11 and 12 on spectral methods.

1.3.2 Mathematical background

As stated in the preface, my goal is to bring the current state of the art in semisupervised learning within the reach of a student or researcher in computational linguistics who has mastered the standard textbooks, in particular, Manning and Schütze, and has acquired a certain familiarity with machine learning through references in the computational linguistics literature, but does not necessarily have a general background in machine learning. This goal is more than a little quixotic. To do things properly, we should lay a foundation of linear algebra, multivariate calculus, optimization theory, probability and statistics, and even a bit of physics (e.g., simple harmonic motion), and on that build a proper treatment of classification and clustering, before

tackling the actual topic of interest, semisupervised learning. But doing so would involve replicating many volumes of material that has been well covered elsewhere. A reader who has already mastered all the background material just mentioned is in an excellent position to tackle the primary literature on semisupervised learning, and will probably not find this book particularly useful. On the other hand, readers who have not mastered all the necessary background material will rightfully feel daunted by the enormity of the task, and would under most circumstances decide that, however interested they may be in semisupervised learning, the cost of entry is simply too great to pay. Those are the readers for whom this book is intended.

My strategy has been to blaze a long thin trail, filling in just the background that is needed to give a reasonably detailed account of the selected semisupervised learning techniques. Two chapters provide an introduction to machine learning: one on classification (chapter 4) and one on clustering (chapter 7). They do not attempt to give a balanced overview of the field, but only to treat topics specifically needed for semisupervised learning. As for more general mathematical background, I have chosen not to collect it into a single chapter – the result would have been a disconnected collection of topics, and the reasons for their inclusion would only have become clear much later. Instead, these topics have been introduced "just in time." The cost is a rather lengthy run-up to the semisupervised techniques involved, especially SVM-based and spectral methods, but that seemed the lesser of the two evils.

1.3.3 Notation

I have collected here notational conventions that I use that are nonstandard or may not be familiar to all readers.

$[\![\Phi]\!]$	semantic value: 1 if Φ is true and 0 otherwise
$\sum_x [\![x \in S]\!] w(x)$	equivalent to: $\sum_{x \in S} w(x)$
$\|\mathbf{x}\|$	vector norm: $\sqrt{\sum_i x_i^2}$
$\|A\|$	cardinality of a set or absolute value of a number
$p[\phi]$	the expectation of ϕ under distribution p
$\tilde{p}(x)$	empirical distribution: relative frequency in sample
$f(x) = \perp$	$f(x)$ is undefined
$\equiv$	is defined as
$F \Rightarrow y$	rule: if the instance has feature F, predict class y
$x \leftarrow x + 1$	set the value of x (in an algorithm)
$\mathbf{D_x}$	derivative with respect to a vector; see section 5.2.3

Bibliography

[1] Abney, Steven. Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),* pp. 360–367. Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[2] Abney, Steven. Understanding the Yarowsky Algorithm. *Computational Linguistics* 30(3):365–395. 2004.

[3] Agichtein, Eugene, Eleazar Eskin, and Luis Gravano. Combining Strategies for Extracting Relations from Text Collections. Columbia University Technical Report CUCS-006-00. 2000.

[4] Agichtein, E. and L. Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. *The 5th International Conference on Digital Libraries.* Association for Computing Machinery. New York. 2000.

[5] Allan, James. Relevance feedback with too much data. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-1995),* pp. 337–343. Association for Computing Machinery. New York. 1995.

[6] Ando, Rie Kubota. Semantic lexicon construction: Learning from unlabeled data via spectral analysis. *Proceedings of the Conference on Natural Language Learning (CoNLL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[7] Baluja, S. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *Advances in Neural Information Processing Systems 11 (NIPS-1998).* MIT Press. Cambridge, MA. 1999.

[8] Basu, S., A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. *Proceedings of the 19th International Conference on Machine Learning (ICML-2002).* Morgan Kaufmann Publishers. San Francisco, CA. 2002.

[9] Bean, David and Ellen Riloff. Unsupervised learning of contextual role knowledge for coreference resolution. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics*

(HLT/NAACL-2004). Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[10] Becker, S. and G.E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* 355:161–163. 1992.

[11] Bengio, S. and Y. Bengio. An EM Algorithm for Asynchronous Input/Output Hidden Markov Models. *Proceedings of the 3rd International Conference On Neural Information Processing (ICONIP 1996),* pp. 328–334. Springer-Verlag. Berlin. 1996.

[12] Bengio, Y. and F. Gingras. Recurrent Neural Networks for Missing or Asynchronous Data. In M. Mozer, D.S. Touretzky, and M. Perrone (eds.), *Advances in Neural Information Processing Systems 8 (NIPS-1995).* MIT Press. Cambridge, MA. 1996.

[13] Bengio, Yoshua, Olivier Delalleau, and Nicolas Le Roux. Label propagation and quadratic criterion. In Chapelle et al. (eds.), *Semi-Supervised Learning* (Chapter 11). MIT Press. Cambridge, MA. 2006.

[14] Bennett, K.P., and A. Demiriz. Semi-supervised support vector machines. *Advances in Neural Information Processing Systems 10 (NIPS-1997),* pp. 368–374. MIT Press. Cambridge, MA. 1998.

[15] Bennett, K., A. Demiriz, and R. Maclin. Exploiting unlabeled data in ensemble methods. *Proceedings of 8th International Conference on Knowledge Discovery and Data Mining (KDD-2002),* pp. 289–296. Association for Computing Machinery. New York. 2002.

[16] Berland, M., and E. Charniak. Finding parts in very large corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999),* pp. 57–64. Association for Computational Linguistics. East Stroudsburg, PA. 1999.

[17] Bhattacharya, Indrajit, Lise Getoor, and Yoshua Bengio. Unsupervised sense disambiguation using bilingual probabilistic models. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[18] Bikel, Daniel M., Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: A High-Performance Learning Name-finder. *Proceedings of the Fifth Conference on Applied Natural Language Processing,* pp. 194–201. 1997.

[19] Bishop, Christopher M. *Neural Networks for Pattern Recognition.* Clarendon Press. Oxford. 1995.

[20] Blum, Avrim and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. *Proceedings of the 18th International Confer-*

ence on Machine Learning, pp. 19–26. Morgan Kaufmann Publishers. San Francisco, CA. 2001.

[21] Blum, Avrim, and Tom Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT),* pp. 92–100. Morgan Kaufmann Publishers. San Francisco, CA. 1998.

[22] Blum, Avrim, John Lafferty, Rajashekar Reddy, and Mugizi Robert Rwebangira. Semi-supervised learning using randomized mincuts. *Proceedings of the 21st International Conference (ICML-2004).* Association for Computing Machinery. New York. 2004.

[23] Bodenreider, Olivier, Thomas Rindflesch, and Anita Burgun. Unsupervised, corpus-based method for extending a biomedical terminology. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Annual Meeting of the Association for Computational Linguistics (ACL-2002).* Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[24] Borthwick, A., J. Sterling, E. Agichtein, and R. Grishman. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. *Sixth Workshop on Very Large Corpora (VLC-1998).* Association for Computational Linguistics. East Stroudsburg, PA. 1998.

[25] Brill, Eric. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. *Third Workshop on Very Large Corpora (VLC-1995),* pp. 1–13. Association for Computational Linguistics. East Stroudsburg, PA. 1995.

[26] Brin, Sergey. Extracting Patterns and Relations from the World Wide Web. *Proceedings of the WebDB Workshop, 6th International Conference on Extending Database Technology (EDBT-1998).* Springer-Verlag. Berlin. 1998.

[27] Briscoe, Ted and John Carroll. Automatic Extraction of Subcategorization from Corpora. *Proceedings of the Conference on Applied Natural Language Processing (ANLP).* Association for Computational Linguistics. East Stroudsburg, PA. 1997.

[28] Briscoe, Ted and John Carroll. Towards Automatic Extraction of Argument Structure from Corpora. Rank Xerox Research Centre Tech Report MLTT-006. 1994.

[29] Broder, Andrei Z., Robert Krauthgamer, and Michael Mitzenmacher. Improved classification via connectivity information. *Proceedings of the Eleventh Annual Symposium on Discrete Algorithms,* pp. 576–585. Association for Computing Machinery/Society for Industrial and Applied Mathematics. New York. 2000.

[30] Brown, P., V. Della Pietra, P. deSouza, J. Lai, and R. Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4):467–480. 1992.

[31] Bruce, Rebecca. "A Statistical Method for Word-Sense Disambiguation." PhD diss. New Mexico State University. 1995.

[32] Buckley, C., M. Mitra, J. Walz, and C. Cardie. Using clustering and SuperConcepts within SMART. *Proceedings of the Sixth Text Retrieval Conference (TREC).* National Institute of Standards and Technology. Gaithersburg, MD. 1998.

[33] Buckley, C., G. Salton, and J. Allan. Automatic retrieval with locality information using SMART. *Proceedings of the First Text Retrieval Conference (TREC),* pp. 59–72. National Institute of Standards and Technology. Gaithersburg, MD. 1992.

[34] Cao, Yunbo, Hang Li, and Li Lian. Uncertainty Reduction in Collaborative Bootstrapping: Measure and Algorithm. *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[35] Caraballo, S.A. Automatic construction of a hypernym-labeled noun hierarchy from text. *Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL-1999).* Association for Computational Linguistics. East Stroudsburg, PA. 1999.

[36] Castelli, Vittorio and Thomas Cover. On the exponential value of labeled samples. *Pattern Recognition Letters* 16:105–111. 1995.

[37] Castelli, V. and T. Cover. The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter. *IEEE Transactions on Information Theory* 42(6):2102–2117. Institute of Electrical and Electronics Engineers. New York. 1996.

[38] Chakrabarti, Soumen. *Mining the Web: Analysis of Hypertext and Semi Structured Data.* Morgan Kaufmann Publishers. San Francisco, CA. 2002.

[39] Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien (eds). *Semi-Supervised Learning.* MIT Press, Cambridge, MA. 2006.

[40] Chapelle, O., J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. *Advances in Neural Information Processing Systems 15 (NIPS-2002).* MIT Press. Cambridge, MA. 2003.

[41] Charniak, Eugene. Unsupervised learning of name structure from coreference data. *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001),* pp. 48–54. Association for Computational Linguistics. East Stroudsburg, PA. 2001.

[42] Chawla, Nitesh and Grigoris Karakoulas. Learning from labeled and unlabeled data: an empirical study across techniques and domains. *Journal of Artificial Intelligence Research* 23:331–366. 2005.

[43] Chelba, Ciprian and Frederick Jelinek. Exploiting Syntactic Structure for Language Modeling. *Proceedings of the 36th Meeting of the Association for Computational Linguistics (ACL-1998).* Association for Computational Linguistics. East Stroudsburg, PA. 1998.

[44] Church, Kenneth. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Texts. *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP-1988).* Association for Computational Linguistics. East Stroudsburg, PA. 1988.

[45] Church, Kenneth, Patrick Hanks, Donald Hindle, William Gale, and Rosamond Moon. Substitutability. Internal Technical Memorandum. AT&T Bell Laboratories. 1990.

[46] Coates-Stephens, Sam. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities* 26:441–456. 1993.

[47] Cohen, I., N. Sebe, F.G. Cozman, and T.S. Huang. Semi-supervised Learning for Facial Expression Recognition. *Proceedings of the 5th International Workshop on Multimedia Information Retrieval (MIR-2003),* pp. 17–22. Association for Computing Machinery. New York. 2003.

[48] Cohen, I., N. Sebe, F.G. Cozman, M.C. Cirelo, and T.S. Huang. Semi-supervised Learning of Classifiers: Theory and Algorithms for Bayesian Network Classifiers and Applications to Human-Computer Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Institute of Electrical and Electronics Engineers. New York. 2004.

[49] Cohn, D.A., Z. Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research* 4:129–145. 1996.

[50] Collins, Michael and Yoram Singer. Unsupervised Models for Named Entity Classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-1999),* pp. 100–110. Association for Computational Linguistics. East Stroudsburg, PA. 1999.

[51] Cormen, T., C. Leiserson, and R. Rivest. *Introduction to Algorithms.* MIT Press. Cambridge, MA. 1990.

[52] Cover, T. and J. Thomas. *Elements of Information Theory.* John Wiley & Sons. New York. 1991.

[53] Cozman, Fabio G. and Ira Cohen. Unlabeled data can degrade classification performance of generative classifiers. *Proceedings of the 15th*

International Florida Artificial Intelligence Research Society Conference (FLAIRS-2002), pp. 327–331. AAAI Press/MIT Press. Cambridge, MA. 2002.

[54] Cozman, Fabio G., I. Cohen, and M.C. Cirelo. Semi-supervised learning of mixture models. *Proceedings of the 20th International Conference on Machine Learning (ICML-2003).* AAAI Press/MIT Press. Cambridge, MA. 2003.

[55] Craven, M. and S. Slattery. Relational Learning with Statistical Predicate Invention: Better Models for Hypertext. *Machine Learning* 43:97–119. 2001.

[56] Craven, Mark, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Sean Slattery. Learning to Construct Knowledge Bases form the World Wide Web. *Artificial Intelligence* 118:69–113. 2000.

[57] Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-1998).* AAAI Press/MIT Press. Cambridge, MA. 1998.

[58] Cucerzan, Silviu and David Yarowsky. Language independent minimally supervised induction of lexical probabilities. *Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL-2000),* pp. 270–277. Association for Computational Linguistics. East Stroudsburg, PA. 2000.

[59] Cucerzan, Silviu and David Yarowsky. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-1999),* pp. 90–99. Association for Computational Linguistics. East Stroudsburg, PA. 1999.

[60] Cucerzan, Silviu and David Yarowsky. Minimally supervised induction of grammatical gender. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[61] Curran, James R. Supersense tagging of unknown nouns using semantic similarity. *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL-2005).* Association for Computational Linguistics. East Stroudsburg, PA. 2005.

[62] Cutting, Doug, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A Practical Part-of-Speech Tagger. *Third Conference on Applied Natural*

Language Processing (ANLP), pp. 133–140. Association for Computational Linguistics. East Stroudsburg, PA. 1992.

[63] Dagan, Ido and Alon Itai. Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics* 20:563–596. 1994.

[64] Dagan, Ido, Alon Itai, and Ulrike Schwall. Two Languages are more Informative than One. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-1991),* pp. 130–137. Association for Computational Linguistics. East Stroudsburg, PA. 1991.

[65] Dagan, Ido, Lillian Lee, and Fernando Pereira. Similarity-based models of word coocurrence probabilities. *Machine Learning* 34(1–3):43–69. 1999.

[66] Dasgupta, Sanjoy. Learning Mixtures of Gaussians. Berkeley Tech Report CSD-99-1047. University of California Berkeley. 1999.

[67] Dasgupta, Sanjoy, Michael Littman, and David McAllester. PAC generalization bounds for co-training. *Advances in Neural Information Processing Systems 14 (NIPS-2001).* MIT Press. Cambridge, MA. 2002.

[68] De Comité, F., F. Denis, R. Gilleron, and F. Letouzey. Positive and unlabeled data help learning. *Proceedings of the Conference on Computational Learning Theory (COLT-1999).* Association for Computing Machinery. New York. 1999.

[69] De Marcken, Carl. The Unsupervised Acquisition of a Lexicon from Continuous Speech. A.I. Memo 1558. MIT AI Lab and Department of Brain and Cognitive Sciences. Massachusetts Institute of Technology. 1995.

[70] De Marcken, Carl. "Unsupervised Language Acquisition." PhD diss. Massachusetts Institute of Technology. 1996.

[71] De Sa, Virginia. Learning classification with unlabeled data. In Cowan, Tesauro, and Alspector (eds), *Advances in Neural Information Processing Systems 6.* Morgan Kaufmann Publishers. San Francisco, CA. 1993.

[72] De Sa, Virginia. "Unsupervised Classification Learning from Cross-Modal Environmental Structure." PhD diss. University of Rochester. 1994.

[73] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic indexing. *Journal of the American Society for Information Science* 41(6):391–407. 1990.

[74] Dempster, A.P., N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38. 1977.

[75] DeRose, S. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics* 14(1). 1988.

[76] Dietterich, Thomas G., and Ghulum Bakiri. Solving multiclass learning problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research* 2:263–286. 1995.

[77] Dietterich, Thomas G., Richard H. Lathrop, and Tomas Lozano-Perez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence* 89:1–2, 31–71. 1997.

[78] Ding, Chris. A tutorial on spectral clustering. *Proceedings of the 21st International Conference on Machine Learning (ICML-2004).* Association for Computing Machinery. New York. 2004.

[79] Doyle, P., and J. Snell. *Random walks and electric networks.* Mathematical Association of America. Washington, D.C. 1984.

[80] Du Toit, S.H.C., A.G.W. Steyn, and R.H. Stumpf. *Graphical Exploratory Data Analysis.* Springer-Verlag. Berlin. 1986.

[81] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification* (second edition). John Wiley & Sons. New York. 2001.

[82] Efthimiadis, N.E. Query expansion. *Annual Review of Information Systems and Technology* 31:121–187. 1996.

[83] Elworthy, David. Does Baum-Welch Re-estimation Help Taggers? *4th Conference on Applied Natural Language Processing (ANLP),* pp. 53–58. Association for Computational Linguistics. East Stroudsburg, PA. 1994.

[84] Etzioni, Oren, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in KnowItAll. *Proceedings of the 13th International Conference on World Wide Web.* Association for Computing Machinery. New York. 2004.

[85] Everitt, B.S. *An Introduction to Latent Variable Models.* Chapman and Hall/Kluwer Academic Publishers. Dordrecht. 1984.

[86] Fawcett, Tom. ROC Graphs: Notes and practical considerations for researchers. HP Labs Tech Report HPL-2003-4. 2003.

[87] Fellbaum, Christiane (ed). *WordNet: An Electronic Lexical Database.* MIT Press. Cambridge, MA. 1998.

[88] Finch, Steven Paul. *Finding Structure in Language.* University of Edinburgh. Edinburgh. 1993.

[89] Flake, Gary W., Kostas Tsioutsiouliklis, and Robert E. Tarjan. Graph clustering techniques based on minimum cut trees. Technical Report 2002-06. NEC Laboratories America. Princeton, NJ. 2002.

[90] Freund, Yoav, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective Sampling Using the Query by Committee Algorithm. *Machine Learning* 28:133–168. 1997.

[91] Gale, W., K. Church, and D. Yarowsky. One sense per discourse. *Proceedings of the Fourth Speech and Natural Language Workshop,* pp. 233–237. Defense Advanced Projects Research Agency (DARPA). Morgan Kaufmann Publishers. San Francisco, CA. 1992.

[92] Ganesalingam, S. and G. McLachlan. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* 65:658–662. 1978.

[93] Ganesalingam, S. and G. McLachlan. Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation* 9:151–158. 1979.

[94] Ghahramani, Zoubin and Michael Jordan. Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems 6 (NIPS-1993).* MIT Press. Cambridge, MA. 1994.

[95] Goldman, Sally and Yan Zhou. Enhancing supervised learning with unlabeled data. *Proceedings of the 17th International Conference on Machine Learning (ICML-2000).* Morgan Kaufmann Publishers. San Francisco, CA. 2000.

[96] Goldsmith, John. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2):153–198. 2001.

[97] Golub, G.H. and C.F.V. Loan. *Matrix Computations* (3rd edition). Johns Hopkins University Press. Baltimore, MD. 1996.

[98] Greenspan, H., R. Goodman, and R. Chellappa. Texture analysis via unsupervised and supervised learning. *International Joint Conference on Neural Networks (IJCNN-1991),* vol. 1, pp. 639–644. Institute of Electrical and Electronics Engineers. New York. 1991.

[99] Grefenstette, G. A new knowledge-poor technique for knowledge extraction from large corpora. *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval (SIGIR-1992).* Association for Computing Machinery. New York. 1992.

[100] Grefenstette, G. Sextant: Extracting semantics from raw text implementation details. Technical Report CS92-05. University of Pittsburgh, Computer Science Dept. 1992.

[101] Grefenstette, G. and M. Hearst. A Knowledge-Poor Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results. *AAAI Workshop on Statistically-Based*

NLP Techniques, 10th National Conference on Artificial Intelligence (AAAI-1992). AAAI Press/MIT Press. Cambridge, MA. 1992.

[102] Hagen, L. and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on CAD* 11:1074–1085. Institute of Electrical and Electronics Engineers. New York. 1992.

[103] Hartigan, J.A. *Clustering Algorithms.* John Wiley & Sons. New York. 1975.

[104] Hartley, H.O. and J.N.K. Rao. Classification and estimation in analysis of variance problems. *Review of International Statistical Institute* 36:141–147. August 1968.

[105] Hastie, T. and W. Stuetzle. Principal curves. *Journal of the American Statistical Association* 84:502–516. 1989.

[106] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer-Verlag. Berlin. 2001.

[107] Hearst, Marti A. Automated discovery of WordNet relations. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database and Some of its Applications.* MIT Press. Cambridge, MA. 1998.

[108] Hearst, M. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics (COLING),* pp. 539–545. Morgan Kaufmann Publishers. San Francisco, CA. 1992.

[109] Hearst, M. Noun homonym disambiguation using local context in large text corpora. *Proceedings, Seventh Annual Conference of the UW Centre for the New OED and Text Research,* pp. 1–22. University of Waterloo. Waterloo, Ontario. 1991.

[110] Hendrickson, Bruce and Robert Leland. A multi-level algorithm for partitioning graphs. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing-1995).* Association for Computing Machinery. New York. 1995.

[111] Hindle, Don and Mats Rooth. Structural ambiguity and lexical relations. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-1991).* Association for Computational Linguistics. East Stroudsburg, PA. 1991.

[112] Hofmann, Thomas. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1–2):177–196. 2001.

[113] Horn, R. and C.R. Johnson. *Matrix Analysis.* Cambridge University Press. Cambridge. 1985.

[114] Inoue, N. Automatic noun classification by using Japanese-English word pairs. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-1991),* pp. 201–208. Association for Computational Linguistics. East Stroudsburg, PA. 1991.

[115] Jaakkola, T., M. Meila, and T. Jebara. Maximum entropy discrimination. *Advances in Neural Information Processing Systems 12 (NIPS-1999).* MIT Press. Cambridge, MA. 2000.

[116] Joachims, Thorsten. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *Proceedings of the 14th International Conference on Machine Learning (ICML-1997).* Morgan Kaufmann Publishers. San Francisco, CA. 1997.

[117] Joachims, T. Transductive inference for text classification using support vector machines. *Proceedings of the 16th International Conference on Machine Learning (ICML-1999).* Morgan Kaufmann Publishers. San Francisco, CA. 1999.

[118] Jones, Rosie, Andrew McCallum, Kamal Nigam, and Ellen Riloff. Bootstrapping for Text Learning Tasks. *Workshop on Text Mining: Foundations, Techniques, and Applications, International Joint Conference on Artificial Intelligence (IJCAI-1999).* Morgan Kaufmann Publishers. San Francisco, CA. 1999.

[119] Jurafsky, D., and J. Martin. *Speech and Language Processing.* Prentice Hall. Upper Saddle River, NJ. 2000.

[120] Karger, David. Random sampling in cut, flow, and network design problems. *Proceedings of the ACM Symposium on Theory of Computing (STOC-1994),* pp. 648–657. Association for Computing Machinery. New York. 1994.

[121] Karger, David R. "Random Sampling in Graph Optimization Problems." PhD diss. Stanford University. 1995.

[122] Karov, Yael and Shimon Edelman. Learning similarity-based word sense disambiguation from sparse data. Technical Report CS-TR 96-05. The Weizmann Institute of Science. 1996.

[123] Kaski, Samuel. Discriminative clustering. *Bulletin of the International Statistical Institute, Invited Paper Proceedings of the 54th Session,* vol. 2, pp. 270–273. International Statistical Institute. Voorburg, The Netherlands. 2003.

[124] Kaufmann, L. and P.J. Rousseeuw. *Finding Groups in Data.* John Wiley & Sons. New York. 1990.

[125] Kishida, Kazuaki. Pseudo relevance feedback method based on Taylor expansion of retrieval function in NTCIR-3 patent retrieval task.

Proceedings of the Workshop on Patent Corpus Processing, Conference on Human Language Technologies and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003). Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[126] Klein, Dan and Christopher D. Manning. Corpus-based induction of syntactic structure: models of dependency and constituency. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[127] Kleinberg, Jon M. and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *IEEE Symposium on Foundations of Computer Science,* pp. 14–23. Institute of Electrical and Electronics Engineers. New York. 1999.

[128] Ko, Youngjoong and Jungyun Seo. Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[129] Kohonen, Teuvo. Improved versions of Learning Vector Quantization. *Proceedings of the International Joint Conference on Neural Networks (IJCNN-1990),* vol. I, pp. 545–550. Institute of Electrical and Electronics Engineers. New York. 1990.

[130] Kuhn, Jonas. Experiments in parallel-text based grammar induction. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[131] Kupiec, Julian. Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language* 6. 1992.

[132] Lafferty, J., A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning (ICML-2001).* Morgan Kaufmann Publishers. San Francisco, CA. 2001.

[133] Lapata, Mirella and Frank Keller. The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[134] Lawrence, Neil D., and Michael I. Jordan. Semi-supervised learning via Gaussian processes. *Advances in Neural Information Processing Systems 17 (NIPS-2004).* MIT Press. Cambridge, MA. 2005.

[135] Lee, Lillian Jane. "Similarity-Based Approaches to Natural Language Processing." PhD diss. Harvard University. 1997.

[136] Lewis, D. and J. Catlett. Heterogenous uncertainty sampling for supervised learning. *Proceedings of the 11th International Conference on Machine Learning (ICML-1994).* Morgan Kaufmann Publishers. San Francisco, CA. 1994.

[137] Lewis, David D. and William A. Gale. A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval (SIGIR-1994).* Association for Computing Machinery. New York. 1994.

[138] Li, Cong and Hang Li. Word translation disambiguation using bilingual bootstrapping. *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL-2002).* Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[139] Li, Hang and Cong Li. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics* 30(1):1–22. 2004.

[140] Li, Hang and Naoki Abe. Clustering Words with the MDL Principle. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-1996).* Morgan Kaufmann Publishers. San Francisco, CA. 1996.

[141] Manning, Chris and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press. Cambridge, MA. 1999.

[142] Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19.2:313–330. 1993.

[143] Mark, Kevin, Michael Miller, and Ulf Grenander. Markov Random Field Models for Natural Languages. *Proceedings of the International Symposium on Information Theory.* Institute of Electrical and Electronics Engineers. New York. 1995.

[144] Mark, Kevin, Michael Miller, Ulf Grenander, and Steve Abney. Parameter Estimation for Constrained Context-Free Language Models. *Proceedings of the Fifth Darpa Workshop on Speech and Natural Language.* Morgan Kaufman Publishers. San Mateo, CA. 1992.

[145] Massart, D.L. and L. Kaufman. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis.* John Wiley & Sons. New York. 1983.

[146] McCallum, Andrew Kachites and Kamal Nigam. Employing EM and Pool-Based Active Learning for Text Classification. *Proceedings of the 15th International Conference on Machine Learning (ICML-1998),* pp. 350–358. Morgan Kaufmann Publishers. San Francisco, CA. 1998.

[147] McCallum, Andrew Kachites and Kamal Nigam. Text Classification by Bootstrapping with Keywords, EM and Shrinkage. *Proceedings of the Workshop on Unsupervised Learning, Meeting of the Association for Computational Linguistics (ACL-1999).* Association for Computational Linguistics. East Stroudsburg, PA. 1999.

[148] McCallum, Andrew Kachites, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval Journal* 3:127–163. 2000.

[149] McLachlan, G. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association* 70:365–369. 1975.

[150] McLachlan, G.J. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American Statistical Association* 72:403–406. 1977.

[151] McLachlan, G.J. and K.E. Basford. *Mixture Models.* Marcel Dekker. New York. 1988.

[152] McLachlan, G.J. and S. Ganesalingam. Updating the discriminant function on the basis of unclassified data. *Communications Statistics – Simulation* 11(6):753–767. 1982.

[153] Merialdo, Bernard. Tagging English Text with a Probabilistic Model. *Computational Linguistics* 20(2):155–172. 1994.

[154] Mihalcea, Rada. Co-training and self-training for word sense disambiguation. *Proceedings of the Conference on Natural Language Learning (CoNLL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[155] Mikheev, Andrei. Unsupervised Learning of Word-Category Guessing Rules. *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-1996).* Association for Computational Linguistics. East Stroudsburg, PA. 1996.

[156] Miller, David and Hasan Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems 9 (NIPS-1996)* pp. 571–577. MIT Press. Cambridge, MA. 1997.

[157] Milun, David. Generating Markov Random Field Image Analysis Systems from Examples. CS Tech Report 95-23. State University of New York/Buffalo. 1995.

[158] Mitchell, T.M. The role of unlabeled data in supervised learning. *Proceedings of the 6th International Colloquium on Cognitive Science.* Kluwer Academic Publishers. Dordrecht. 1999.

[159] Morris, Stephen. Contagion. *Review of Economic Studies* 67:57–58. 2000.

[160] Muslea, Ion, Steven Minton, and Craig A. Knoblock. Active + Semi-Supervised Learning = Robust Multi-View Learning. *Proceedings of the 19th International Conference on Machine Learning (ICML-2002).* Morgan Kaufmann Publishers. San Francisco, CA. 2002.

[161] Muslea, Ion, Steven Minton, and Craig A. Knoblock. Selective Sampling With Redundant Views. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000).* AAAI Press/MIT Press. Cambridge, MA. 2000.

[162] Navigli, Roberto and Paola Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics* 30(2). 2004.

[163] Ng, Vincent and Claire Cardie. Weakly supervised natural language learning without redundant views. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[164] Nigam, Kamal. "Using Unlabeled Data to Improve Text Classification." PhD diss. Tech Report CMU-CS-01-126. Computer Science, Carnegie Mellon University. 2001.

[165] Nigam, Kamal and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. *Proceedings of the 9th International Conference on Information and Knowledge Management.* Association for Computing Machinery. New York. 2000.

[166] Nigam, Kamal and Rayid Ghani. Understanding the Behavior of Co-training. *Proceedings of the Workshop on Text Mining, 6th International Conference on Knowledge Discovery and Databases (KDD-2000).* Association for Computing Machinery. New York. 2000.

[167] Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* 39:103–134. 2000.

[168] Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Learning to Classify Text from Labeled and Unlabeled Documents. *Proceedings of the 15th National Conference on Artificial*

Intelligence (AAAI-1998). AAAI Press/MIT Press. Cambridge, MA. 1998.

[169] Niu, Cheng, Wei Li, Jihong Ding, and Rohini K. Srihari. A bootstrapping approach to named entity classification using successive learners. *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[170] Niu, Cheng, Wei Li, and Rohini K. Srihari. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[171] Oflazer, Kemal, Marjorie McShane, and Sergei Nirenburg. Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics* 27(1). 2001.

[172] O'Neill, T.J. Normal discrimination with unclassified observations. *Journal of the American Statistical Association* 73(364):821–826. 1978.

[173] Osborne, Miles and Jason Baldridge. Ensemble-based active learning for parse selection. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[174] Pakhomov, Serguei. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL-2002).* Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[175] Pantel, Patrick and Deepak Ravichandran. Automatically labeling semantic classes. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[176] Pantel, Patrick and Dekang Lin. An unsupervised approach to prepositional phrase attachment using contextually similar words. *Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL-2000).* Association for Computational Linguistics. East Stroudsburg, PA. 2000.

[177] Pao, Y.-H. and D.J. Sobajic. Combined use of supervised and unsupervised learning for dynamic security assessment. *IEEE Transactions on Power Systems* 7(2):878–884. Institute of Electrical and Electronics Engineers. New York. 1992.

[178] Papadimitriou, Christos H., Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences* 61:217–235. 2000.

[179] Peng, Fuchun and Andrew McCallum. Accurate information extraction from research papers using conditional random fields. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[180] Pereira, Fernando and Yves Schabes. Inside-outside reestimation from partially bracketed corpora. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-1992),* pp. 128–135. Association for Computational Linguistics. East Stroudsburg, PA. 1992.

[181] Pereira, Fernando, Naftali Tishby, and Lillian Lee. Distributional Clustering of English Words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-1993),* pp. 183–190. Association for Computational Linguistics. East Stroudsburg, PA. 1993.

[182] Phillips, William, and Ellen Riloff. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).* Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[183] Purandare, Amruta and Ted Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. *Proceedings of the Conference on Natural Language Learning (CoNLL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[184] Quillian, M. Semantic memory. In M. Minsky (ed.), *Semantic Information Processing,* 227–270. MIT Press. Cambridge, MA. 1968.

[185] Ratsaby, Joel. "The Complexity of Learning from a Mixture of Labeled and Unlabeled Examples." PhD diss. University of Pennsylvania. 1994.

[186] Ratsaby, Joel and Santosh S. Vankatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. *Proceedings of 8th Annual Conference on Computational Learning Theory (COLT-1995),* pp. 412–417. Association for Computing Machinery. New York. 1995.

[187] Rattray, Magnus. A model-based distance for clustering. *Proceedings of the International Joint Conference on Neural Networks (IJCNN-2000).* Institute of Electrical and Electronics Engineers. New York. 2000.

[188] Riloff, Ellen. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *AI Journal* 85. August 1996.

[189] Riloff, Ellen. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-1993).* AAAI Press/MIT Press. Cambridge, MA. 1993.

[190] Riloff, Ellen. Automatically Generating Extraction Patterns from Untagged Text. *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996).* AAAI Press/MIT Press. Cambridge, MA. 1996.

[191] Riloff, E. and M. Schmelzenbach. An Empirical Approach to Conceptual Case Frame Acquisition. *Proceedings of the Sixth Workshop on Very Large Corpora (VLC-1998).* Association for Computational Linguistics. East Stroudsburg, PA. 1998.

[192] Riloff, E. and J. Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-1997),* pp. 127–132. Association for Computational Linguistics. East Stroudsburg, PA. 1997.

[193] Riloff, E. and J. Shepherd. A Corpus-Based Bootstrapping Algorithm for Semi-Automated Semantic Lexicon Construction. *Journal of Natural Language Engineering* 5(2):147–156. 1999.

[194] Riloff, Ellen and Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-1999).* AAAI Press/MIT Press. Cambridge, MA. 1999.

[195] Roark, Brian and Eugene Charniak. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics.* Association for Computational Linguistics. East Stroudsburg, PA. 1998.

[196] Rocchio, J.J., Jr. Relevance feedback in information retrieval. In G. Salton (ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall. Englewood Cliffs, NJ. 1971.

[197] Roget, Peter Mark. *Roget's Thesaurus of English Words and Phrases.* 1911 edition available from Project Gutenberg (`http://www.gutenberg.org/`). 2004.

[198] Roth, Dan and Dmitry Zelenko. Toward a Theory of Learning Coherent Concepts. *Proceedings of the 17th National Conference on Artificial*

Intelligence (AAAI-2000). AAAI Press/MIT Press. Cambridge, MA. 2000.

[199] Russell, Stuart J. and Peter Norvig. *Artificial Intelligence: A Modern Approach*, 2nd Edition. Prentice Hall. Upper Saddle River, NJ. 2002.

[200] Samuel, A. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3:210–229. 1959. (Reprinted in vol. 44:206–227, 2000.)

[201] Sapir, Edward. *Language: An Introduction to the Study of Speech.* Harcourt, Brace and Company. New York. 1921.

[202] Sarkar, Anoop. "Combining Labeled and Unlabeled Data in Statistical Natural Language Parsing." PhD diss. University of Pennsylvania. 2001.

[203] Schapire, Robert, Yoram Singer, and Amit Singhal. Boosting and Rocchio applied to text filtering. *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR-1998).* Association for Computing Machinery. New York. 1998.

[204] Schütze, Hinrich. Word Space. *Advances in Neural Information Processing Systems* 5. Morgan Kaufmann Publishers. San Mateo, CA. 1993.

[205] Sedgewick, Robert. *Algorithms.* Second edition. Addison-Wesley. Reading, MA. 1988.

[206] Seeger, M. Learning with labeled and unlabeled data. Technical Report, University of Edinburgh. Edinburgh. 2001.

[207] Shahshahani, B.M. and D.A. Landgrebe. On the asymptotic improvement of supervised learning by utilizing additional unlabeled samples; normal mixture density case. *Proceedings of the International Conference on Neural and Stochastic Methods in Image and Signal Processing* 1766:143–155. Society of Photographic Instrumentation Engineers (SPIE). Bellingham, WA. 1992.

[208] Shahshahani, B.M. and D.A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing* 32(5):1087–1095. Institute of Electrical and Electronics Engineers. New York. 1994.

[209] Shen, Dan, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. Multi-criteria-based active learning for named entity recognition. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[210] Sinkkonen, Janne and Samuel Kaski. Semisupervised clustering based on conditional distributions in an auxiliary space. Technical Report A60. Helsinki University of Technology, Publications in Computer and Information Science. Espoo, Finland. 2000.

[211] Smith, Noah A. and Jason Eisner. Annealing techniques for unsupervised statistical language learning. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[212] Steedman, Mark, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. Example selection for bootstrapping statistical parsers. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[213] Strang, Gilbert. *Introduction to Linear Algebra.* Third edition. Wellesley-Cambridge Press. Wellesley, MA. 2003.

[214] Sudo, Kiyoshi, Satoshi Sekine, and Ralph Grishman. An improved extraction pattern representation model for automatic IE pattern acquisition. *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[215] Szummer, M. and T. Jaakkola. Kernel expansions with unlabeled examples. *Advances in Neural Information Processing Systems 13 (NIPS-2000).* MIT Press. Cambridge, MA. 2001.

[216] Szummer, M. and T. Jaakkola. Partially labeled classification with Markov random walks. *Advances in Neural Information Processing Systems 14 (NIPS-2001).* MIT Press. Cambridge, MA. 2002.

[217] Tang, Min, Xiaoqiang Luo, and Salim Roukos. Active learning for statistical natural language parsing. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),* pp. 120–127. Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[218] Tesauro, Gerald. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation* 6.2:215–219. 1994.

[219] Thelen, Michael, and Ellen Riloff. A bootstrapping method for learning semantic lexicons using extracting pattern contexts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).* Association for Computational Linguistics. East Stroudsburg, PA. 2002.

[220] Thrun, S. Exploration in Active Learning. In M. Arbib (ed.), *Handbook of Brain and Cognitive Science.* MIT Press. Cambridge, MA. 1995.

[221] Tipping, M. Deriving cluster analytic distance functions from Gaussian mixture models. *Proceedings of the 9th International Conference on Artificial Neural Networks.* Institute of Electrical and Electronics Engineers. New York. 1999.

[222] Tishby, Naftali Z., Fernando Pereira, and William Bialek. The information bottleneck method. In Bruce Hajek and R.S. Sreenivas, eds., *Proceedings of the 37th Allerton Conference on Communication, Control and Computing.* University of Illinois/Urbana-Champaign. Urbana, Illinois. 1999.

[223] Titterington, D.M., A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons. New York. 1985.

[224] Tolat, V.V. and A.M. Peterson. Nonlinear mapping with minimal supervised learning. *Proceedings of the Hawaii International Conference on System Science* 1. Jan 1990.

[225] Tolstov, Georgi P. *Fourier Series.* Prentice-Hall. Englewood Cliffs, NJ. 1962.

[226] Tong, S. and D. Koller. Support vector machine active learning with applications to text classification. *Proceedings of the 17th International Conference on Machine Learning (ICML-2000),* pp. 999–1006. Morgan Kaufmann Publishers. San Francisco, CA. 2000.

[227] Vapnik, Vladimir N. *Statistical Learning Theory.* John Wiley & Sons. New York. 1998.

[228] Weiss, G.M. and F. Provost. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research* 19:315–354. 2003.

[229] Weston, Jason et al. Protein ranking: From local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences of the United States of America* 101(17):6559–6563. 2004.

[230] Widdows, Dominic. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *Proceedings of the Conference on Human Language Technology and the Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[231] Widdows, Dominic, and Beate Dorow. A graph model for unsupervised lexical acquisition. *19th International Conference on Computational Linguistics (COLING-2002),* pp. 1093–1099. Morgan Kaufmann Publishers. San Francisco, CA. 2002.

[232] Widdows, Dominic, Stanley Peters, Scott Cederberg, Chiu-Ki Chan, Diana Steffen, and Paul Buitelaar. Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS. *Workshop on Natural Language Processing in Biomedicine, 41st Meeting of the Association for Computational Linguistics (ACL-2003),* pp. 9–16. Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[233] Winkler, Gerhard. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer-Verlag. Berlin. 1995.

[234] *Wordnet Reference Manual.* `http://wordnet.princeton.edu/doc`.

[235] Wu, Dekai, Weifeng Su, and Marine Carpuat. A kernel PCA method for superior word sense disambiguation. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-2004).* Association for Computational Linguistics. East Stroudsburg, PA. 2004.

[236] Yangarber, Roman. Counter-training in discovery of semantic patterns. *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL-2003).* Association for Computational Linguistics. East Stroudsburg, PA. 2003.

[237] Yangarber, Roman, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. *Proceedings of the Conference on Applied Natural Language Processing and the Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000),* pp. 282–289. Association for Computational Linguistics. East Stroudsburg, PA. 2000.

[238] Yarowsky, David. One sense per collocation. *Proceedings of the Advanced Research Projects Agency (ARPA) Workshop on Human Language Technology.* Morgan Kaufmann Publishers. San Mateo, CA. 1993.

[239] Yarowsky, David. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995),* pp. 189–196. Association for Computational Linguistics. East Stroudsburg, PA. 1995.

[240] Yarowsky, D. and R. Wicentowski. Minimally supervised morphological analysis by multimodal alignment. *Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL-2000),* pp. 207–216. Association for Computational Linguistics. East Stroudsburg, PA. 2000.

[241] Zhang, Q., and Sally Goldman. EM-DD: An Improved Multiple-Instance Learning Technique. *Advances in Neural Information Processing Systems 14 (NIPS-2001).* MIT Press. Cambridge, MA. 2002.

[242] Zhang, Tong. The value of unlabeled data for classification problems. *Proceedings of the 17th International Conference on Machine Learning (ICML-2000).* Morgan Kaufmann Publishers. San Francisco, CA. 2000.

[243] Zhang, Tong and Frank J. Oles. A probability analysis on the value of unlabeled data for classification problems. *Proceedings of the 17th International Conference on Machine Learning (ICML-2000),* pp. 1191–1198. Morgan Kaufmann Publishers. San Francisco, CA. 2000.

[244] Zhu, Xiaojin. "Semi-Supervised Learning with Graphs." PhD diss. Carnegie Mellon University. 2005.

[245] Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. *Proceedings of the 20th International Conference on Machine Learning (ICML-2003).* AAAI Press/MIT Press. Cambridge, MA. 2003.

[246] Zhu, Xiaojin, John Laffery, and Zoubin Ghahramani. Semi-supervised learning: from Gaussian fields to Gaussian processes. CMU Tech Report CMU-CS-03-175. Carnegie Mellon University. 2003.

[247] Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *Proceedings of the 20th International Conference on Machine Learning (ICML-2003).* AAAI Press/MIT Press. Cambridge, MA. 2003.