Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing

Stefan Wermter
Ellen Riloff
Gabriele Scheler

Springer, Heidelberg, New York

March, 1996

Preface

The purpose of this book is to present a collection of papers that represents a broad spectrum of current research in learning methods for natural language processing, and to advance the state of the art in language learning and artificial intelligence. The book should bridge a gap between several areas that are usually discussed separately, including connectionist, statistical, and symbolic methods.

In order to bring together new and different language learning approaches, we held a workshop at the International Joint Conference on Artificial Intelligence in Montreal in August 1995. Paper contributions were selected and revised after having been reviewed by at least two members of the international program committee as well as additional reviewers. This book contains the revised workshop papers and additional papers by members of the program committee.

In particular this book focuses on current issues such as:

- How can we apply existing learning methods to language processing?
- What new learning methods are needed for language processing and why?
- What language knowledge should be learned and why?
- What are the similarities and differences between different approaches?
- What are the strengths of learning as opposed to manual encoding?
- How can learning and manual encoding be combined?
- Which aspects of system architectures have to be considered?
- What are successful applications of learning methods in various fields?
- How can we evaluate learning methods using real-world language?

We believe that this selection of contributions is a representative snapshot of the state of the art in current approaches to learning for natural language processing. This is an extremely active area of research that is growing rapidly in interest and popularity. Systems built by learning methods have reached a level where they can be applied to real-world problems in natural language processing and where they can be compared with more traditional encoding methods. The book will provide a basis for discussing various learning approaches to support natural language processing. We hope that this collection will be stimulating and useful for all interested in the areas of learning and natural language processing.

January 1996

Stefan Wermter
Ellen Riloff
Gabriele Scheler

ORGANIZING COMMITTEE

Stefan Wermter, University of Hamburg, Germany
Ellen Riloff, University of Utah, USA
Gabriele Scheler, Technical University Munich, Germany

INVITED SPEAKERS

Eugene Charniak, Brown University, USA
Noel Sharkey, Sheffield University, UK

PROGRAM COMMITTEE

Jaime Carbonell, Carnegie Mellon University, USA
Joachim Diederich, Queensland University of Technology, Australia
Georg Dorffner, University of Vienna, Austria
Jerry Feldman, ICSI, Berkeley, USA
Walther von Hahn, University of Hamburg, Germany
Aravind Joshi, University of Pennsylvania, USA
Ellen Riloff, University of Utah, USA
Gabriele Scheler, Technical University Munich, Germany
Stefan Wermter, University of Hamburg, Germany

ACKNOWLEDGMENTS

We would like to thank Katja Hillebrand, Uwe Jost, Alexandra Klein, Eva Landmann, Manuela Meurer, Jens-Uwe Möller, and Volker Weber from the University of Hamburg for their important help and assistance during the preparations of the workshop and the book. Furthermore, we would like to thank Tony Cohn, the IJCAI-95 workshop chair, and Carol Hamilton at AAAI for their cooperation. We would like to thank Alfred Hofmann and Anna Kramer from Springer for the effective cooperation. Finally, we thank the participants of the workshop and the contributors to this book.

Table of Contents

Statistical Approaches

Symbolic Approaches

Learning Approaches for Natural Language Processing

Stefan Wermter[1] and Ellen Riloff[2] and Gabriele Scheler[3]

[1] University of Hamburg, Dept. of Computer Science, 22527 Hamburg, Germany
[2] University of Utah, Dept. of Computer Science, Salt Lake City, UT 84112, USA
[3] Technical University Munich, Dept. of Computer Science, 80290 Munich, Germany

Abstract. The purpose of this chapter is to provide an introduction to the field of connectionist, statistical and symbolic approaches to learning for natural language processing, based on the contributions in this book. The introduction has been split into three parts: (1) neural networks and connectionist approaches, (2) statistical approaches, and (3) symbolic machine learning approaches. We will give a brief overview of the main methods used in each field, summarize the work that is presented here, and provide some additional references. In the final section we will highlight important general issues and trends based on the workshop discussions and book contributions.

1 Introduction

In the last few years, there has been a lot of interest and activity in developing new approaches to learning for natural language processing [66, 52, 80, 24, 11, 39]. Various learning methods have been used, including connectionist methods/neural networks, statistical methods, symbolic machine learning algorithms, genetic methods, and various hybrid approaches.

In general, learning methods are designed to support automated knowledge acquisition, fault tolerance, and plausible induction. Using learning methods for natural language processing is especially important because learning is an enabling technology for many language-related tasks, such as speech recognition, spoken language understanding, machine translation, and information retrieval. Furthermore, learning is important for building more flexible, scalable, adaptable, and portable natural language systems.

A complete survey of all research in the field of connectionist, statistical, and symbolic approaches to learning for natural language processing cannot be our goal in this chapter. The field has grown much too large already. In fact, such a survey could easily fill a textbook of its own. However, as an aid to the reader, we will set the stage for the contributions in this book. We have organized the book and this survey chapter into three parts for connectionist, statistical, and symbolic learning approaches. In each of the three parts of this chapter, we will give (1) a brief introduction to the general approach, (2) a description, categorization and discussion of the paper contributions, and (3) some pointers to further work.

2 Connectionist Networks and Hybrid Approaches

2.1 Introduction

Recently, artificial neural networks and connectionist networks[4] have received a lot of attention as computational learning mechanisms for natural language processing [66, 52, 24, 2, 80]. There exist many different architectures for connectionist networks. For instance, feedforward networks are useful because of their universal function approximation qualities [34, 16]. While feedforward networks can represent only a fixed length input, recurrent networks can represent variable length input [23, 40, 61]. Most connectionist networks can be trained with a specific learning rule based on the network architecture and the units used.

Connectionist networks have been shown to model successfully a whole variety of language learning tasks [60, 72, 45, 32]. In addition, the combination or integration of connectionist networks with statistical and symbolic representations is an important field for natural language processing [33, 22, 79, 76, 80]. From the viewpoint of knowledge engineering, it might be efficient to encode well-known rules rather than learning them from scratch. From the viewpoint of cognitive behavior, it is interesting to explore human symbolic reasoning for natural language processing based on neural architectures. We speak of a hybrid connectionist approach either if several different connectionist networks are integrated or if connectionist networks are integrated with symbolic or statistical methods.

2.2 Connectionist Approaches for Syntax and Semantics

In this book, the work described by Sharkey and Sharkey argues for a separation of learning and representation. They trained a simple recurrent network with sequences generated by a finite state grammar with bidirectional links. An analysis showed that the networks had a restricted capability to encode embeddings. Then a constructive method was used to improve the performance of the network. Incorrect predictions were identified and hidden units responsible for these predictions were found. A new retraining phase for these incorrect predictions followed and led to better prediction performance. This work might lead to the incremental development of special neural network architectures which cannot be built easily without constructive methods.

The underlying motivation of the approach taken by Lawrence, Fong, and Giles is to explore whether a neural network can exhibit the same discriminatory power for grammaticality as linguists have claimed to exist in universal principles and learned parameters. They have compared different neural network algorithms and symbolic machine learning algorithms. Recurrent Elman networks provided the best discriminatory performance on training and test sets.

[4] The terms artificial neural networks and connectionist networks are often used in a similar manner, both refer to networks which are based on very rough computational models of biological neurons.

The authors claim that the Elman network has learned a state machine which can discriminate between grammatical and ungrammatical sentences while other learning algorithms like feedforward models, decision trees, and nearest neighbor algorithms have only learned to find closest matches.

Hayward, Tickle, and Diederich focus on a detailed analysis of rule representation in a single connectionist network. A network is trained with simple sentences using an extension of the cascade correlation algorithm. The task is to learn which combinations of noun – verb – noun combinations are grammatical. Usually little emphasis is put on extracting the rules from a network. In this approach it is demonstrated that simple rules can be extracted from a Cascade network. While the training corpus is relatively simple, this is one of the few papers which focuses on the explanation of connectionist representations and the possible extraction of rules.

The work presented by Scheler uses supervised learning for the construction of classification functions from semantic representations to overt grammatical categories, and interpretation functions from texts to semantic representations. Semantic representations consist of a set of atomic, logically interpretable features which are grounded in cognitive representations. They can be derived automatically from surface coding of texts with sufficient accuracy to provide grammar checking for definiteness of English noun phrases. The results are considerably better than a "naive" approach which classifies surface encodings directly. With this work the difficult issue of providing logical interpretations for real texts automatically has been confronted and a general solution using connectionist learning methods has been presented.

Winiwarter, Schweighofer, and Merkl explore the use of unsupervised learning techniques for knowledge acquisition in concept and document spaces. In particular unsupervised learning in Kohonen feature maps is adapted to cluster legal text segments. This unsupervised learning method was applied to cluster full text documents of court decisions and it could be shown that most similar documents fell into similar regions.

2.3 Hybrid Approaches for Spoken Language

Spontaneously spoken language can be very erroneous. At the acoustic level, phonemes or words may be incorrectly analyzed, and restarts, interjections, pauses, repairs, and repetitions often occur. Furthermore, sentences may be grammatically or semantically incorrect. In general, it is hardly possible to encode all necessary knowledge in fault-tolerant rules. Connectionist networks have been examined for spoken language analysis due to their support of learning and fault-tolerance. The integration of connectionist approaches with symbolic approaches has also been explored.

The approach described by Weber and Wermter tackles the analysis of spoken language in the hybrid connectionist architecture SCREEN. Based on a speech recognizer, spontaneously spoken sentences are processed. Feedforward and recurrent connectionist networks for semantic and syntactic analysis are used wherever possible, but symbolic techniques are also used in a restricted manner for the control of different networks and for simple rules which always hold. The focus in this chapter is on improving the analysis of noisy and ungrammatical spoken sentences by integrating acoustic, syntactic, and semantic knowledge. Furthermore, it is shown that a flat analysis using connectionist learning supports a robust and fault-tolerant analysis of spoken language.

The hybrid symbolic/connectionist architecture SKOPE for spoken Korean by Lee and Lee uses connectionist learning in Time Delay networks, primarily in the speech component. Time-delay networks offer a method of learning a sequence of events. The morphological and syntactical components are based on table-driven parsing techniques and spreading activation respectively. In general, Lee and Lee argue for the combination of connectionist speech learning with symbolic language encoding.

Geutner and colleagues combine different learning approaches for translating spoken natural language between English, German, and Spanish[5]. Within their JANUS system, statistical hidden Markov models and connectionist Time Delay Networks are explored at the speech recognition level. Subsequently, a concept spotter, a statistical LR parser and a connectionist parser are examined to provide an interlingua-like language description for the translation. A main point within the JANUS system is that multiple learning strategies are explored in a complementary manner to make hand-coded language representations unnecessary.

2.4 Genetic Approaches

The paper by Smith and Witten describes a genetic algorithm for the induction of natural language grammars. The genetic learning algorithm works on logical s-expressions which can represent context free grammars without recursion. Starting with a random population of different potential grammars for a given first string, only those which parse the string are part of the initial population. If a new string is added, reproduction operators combine two different grammars from the current population. Furthermore, additional mutation operators can replace leaf or internal nodes in a grammar in order to provide more variation in the search space of possible grammars. Although relatively few, small sentences are used for genetic learning, this approach is a new interesting alternative since it is also frequency-based and robust.

2.5 Further Work

There are a large number of important references in the field of connectionist natural language processing and we cannot hope to be complete or comprehensive. However, in order to provide some pointers to start with, we will very briefly list some of the important references. For a fundamental early

[5] This paper could also be grouped into the statistical section. However, because of the many different learning strategies used and due to the focus on spoken language it was placed in the group of hybrid approaches for spoken language.

paper about connectionism in general one could start with [25, 51]. Architectural issues of connectionist and hybrid connectionist systems are discussed in [68, 22, 23, 40, 61, 3, 21, 52, 19, 56, 80]. Some representative references for semantic and syntactic analysis with connectionist networks can be found in [38, 50, 60, 75, 70, 79]. For references on cognitively oriented connectionist natural language processing some references are [14, 78, 69, 42, 12].

3 Statistical Approaches

3.1 Introduction

With the recent trend for learning in natural language processing, statistical methods have gained new popularity, and are being applied to new domains. They are usually characterized by using large text corpora and performing some analysis which uses primarily the text characteristics without adding significant linguistic or world knowledge [5, 11, 48]. Text corpora that have been built include the still widely used Brown corpus [27], and newer corpora such as the LOB corpus [28] and the Penn treebank [49].

Annotation of corpora with part-of-speech tags or parse trees has been a focus of corpus-based language analysis. Additional important application areas of statistical techniques to written natural language are thesaurus-building (or lexical clustering) and probabilistic grammar learning. Statistical techniques that have been used for these tasks are n-gram techniques, unsupervised clustering and hidden Markov models (e.g. [71, 44]). A special case is grammar induction, which uses context-free grammars in addition to probabilistic information from texts.

3.2 Probabilistic Grammar Induction and Disambiguation

In this book, Ersan and Charniak present a system for syntactic disambiguation using probabilistic information on word classes derived from the Wall Street Journal corpus. They show how the improved parser can be used to extract data on verb case-frames and noun-preposition and adjective-preposition combinations. This is achieved by identifying occurrences of particular syntactic combinations in the parses, which are counted and passed through a probabilistic filter. Precision and recall for the resulting combinations are evaluated with respect to an English dictionary. This statistical work is particularly interesting since it addresses the important question what is learned in probabilistic language representations.

Hogenhout and Matsumoto also collect general statistics on the occurrence of semantic classes in the application of grammar rules. That is, the syntactic word class information is augmented by the probability of a semantic class in the application of a single rule within a context-free grammar. This probability is calculated by the Inside-Out algorithm with certain smoothing parameters.

Semantic classes are defined by reference to a standard thesaurus. The probabilities of the classes were used to improve ambiguity resolution within a handwritten grammar. Experiments conducted on the Japanese EDR corpus show a statistically significant improvement on parsing accuracy (sentence and bracket accuracy) for the incorporation of probabilistic semantic class information.

Fong and Wu describe a model for learning probabilistic link grammars. Link grammars are highly-lexicalized context-free grammars where individual words can be linked via labeled arcs. The probabilities of the links are estimated using an expectation maximization training method. After training and subsequent pruning, the learned representation contains grammatical rules as sets of simple disjuncts and probabilities. This approach was tested with two artificial corpora of short simple sentences to demonstrate the learning behavior. It could be shown that the perplexity of the described model is lower than a comparable probabilistic link model as well as a bigram model.

Prepositional phrase attachment has been a major problem for structural analysis of natural language. Franz describes a statistical approach to learning prepositional phrase attachment based on categorial features. A loglinear model is described which consists of a contingency table for recording the frequency of certain feature combinations as well as a loglinear model for smoothing the frequency counts for zero occurrences. The Brown corpus and Wall Street Journal corpus were used for training and testing. The results are slightly worse than human performance but better than simple heuristics like right association. Learning in this statistical model is simple but can be applied efficiently to a large number of training and test instances.

3.3 Part-of-speech Tagging and Probabilistic Word Classes

Grünwald uses a greedy minimum description length (MDL) approach to cluster words in semantic-syntactic classes, based on a subset of the Brown corpus. According to the MDL principle, learning is defined as reducing the total length of a set of data (measured in bits) by introducing a theory which can generate certain data, and thus serves as an abbreviation of the data set. The implementation uses a "greedy" learning mechanism, i.e. decreasing the total description length in each step. As a result, a number of the derived word classes are given. The advantage of the MDL approach in comparison with a simple n-gram technique is the availability of a stopping criterion for learning, which prevents overfitting of the data.

The approach by Mast and colleagues offers a new application area, namely the classification of dialog acts. Dialog acts describe a spoken dialog at a higher level using shallow understanding with labels like "accept", "reject", "request-suggestion". State-dependent semantic classification trees and statistical polygrams are used to acquire a classification of sentences according to their respective dialog acts. German and English dialogs were labeled and used for this task. It is argued that polygrams are preferable to the decision tree methods for dialog act classification.

Selective sampling is a technique for selecting only particularly informative unlabeled training examples for subsequent labeling and training. Engelson and Dagan describe an approach for selective sampling applied to probabilistic part-of-speech tagging. Using an implicit model, the current training data is used to evaluate the uncertainty for classifying an additional training example. Examples with a larger uncertainty for classification are particularly good training examples for labeling and training. Since labeling large corpora is very expensive, time-consuming work, the technique of selective sampling could allow systems to work with much larger corpora.

3.4 Further Work

The classical applications in natural language processing are part-of-speech tagging [46, 15, 18], and lexical extraction for various information retrieval tasks [13, 77, 35, 30, 73]. This has recently been extended to anaphora resolution [8], text alignment [41], grammar induction [6] and statistical machine translation [59, 7]. Statistical analysis has also been used in speech recognition [4], which is however not the main focus of the present volume.

4 Symbolic Approaches

4.1 Introduction

Symbolic approaches to learning encompass a wide variety of machine learning techniques. Many inductive learning algorithms have been developed in the machine learning community, such as decision tree algorithms [63, 65] and conceptual clustering [26]. Explanation-based learning [17, 53] is another type of symbolic learning that pushes training examples through a domain theory to create generalized examples for future use. Case-based learning techniques [31, 62] and analogical reasoning methods [10] try to map new situations onto previously encountered situations to find the best solution. There are also a wide variety of rule-based approaches to concept learning.

Information extraction (IE) is a relatively new subfield of natural language processing that has received a lot of attention recently because of the message understanding conferences (MUCs) [57, 58]. The MUCs have encouraged researchers to work on real-world text (e.g., newswire articles) and to develop practical methods, and have been instrumental in bringing together NLP researchers from a variety of areas towards a common goal. There has also been growing interest in developing trainable IE systems that can use learning methods to increase their portability to new domains; some of these systems are mentioned in the next section.

4.2 Information Extraction

One of the main challenges in information extraction research is developing methods and systems to acquire the necessary knowledge bases automatically. Huffman has developed one such system, LIEP, which learns dictionaries of extraction patterns. LIEP attempts to find relationships between constituents that have been tagged as relevant by a user. One of the distinguishing features of LIEP is that the underlying sentence analyzer, ODIE, parses sentences "on-demand" by attempting to verify syntactic relationships only when asked to do so. LIEP hypothesizes syntactic relationships between constituents and asks ODIE to determine whether the relationships are plausible. This approach chooses among competing patterns using empirical feedback on a training corpus and can generalize existing patterns when the same syntactic relationships are identified in a new context.

Hastings describes the CAMILLE lexical acquisition system which was originally developed to learn word meanings for an information extraction system. CAMILLE infers the meanings of new words based on semantic constraints provided by the surrounding context and a concept hierarchy. This paper explains why different learning strategies are used for nouns and verbs, and discusses implications for related research in knowledge representation, cognitive modeling, and evaluation.

Riloff discusses the application of learned extraction patterns to problems in text classification. The AutoSlog dictionary construction system learns extraction patterns automatically using an annotated training corpus and the learned patterns have been shown to be effective for information extraction. This paper presents experiments in three domains which show that the extraction patterns created by AutoSlog are also useful for text classification. These results demonstrate that AutoSlog's dictionaries represent important domain concepts and that information extraction dictionaries can be useful for other natural language processing tasks as well.

The CRYSTAL system by Soderland and colleagues automatically learns information extraction patterns. CRYSTAL is an automated dictionary construction tool that uses an annotated training corpus and a concept hierarchy to generate extraction patterns. The produced extraction patterns are tested on a training corpus to ensure they satisfy a minimum error tolerance threshold. This paper discusses the issues of creating an appropriate training corpus and domain ontology, the expressiveness of its learned patterns, and its search control strategy.

4.3 Inductive Learning Algorithms

Many researchers are taking advantage of inductive learning methods developed in the machine learning community and are applying them to problems in natural language processing. A good example of this type of work is the application of the C4.5 decision tree algorithm to anaphora resolution by Aone and Bennett. Newspaper articles annotated with discourse information are given to C4.5 as training instances. Aone and Bennet use 66 features to represent each training instance; the feature values are determined automatically during text processing. Their paper presents a series of experiments using different parameter combinations (such as anaphoric chaining, anaphoric type identification, and confidence factors) in the context of a full information extraction system, and compares the performance of the learning system with hand-coded knowledge sources.

Cardie promotes the view that all ambiguity problems in NLP can be recast as classification tasks and presents a general architecture for embedding machine learning techniques in natural language processing systems. The Kenmore framework is composed of a text corpus, a sentence analyzer, a human supervisor, and a machine learning algorithm. In the acquisition phase, texts are parsed by the sentence analyzer to produce training cases, which are then annotated by a human and presented to the machine learning algorithm as training data. Cardie describes experiments with Kenmore for several ambiguity problems using a hybrid nearest-neighbor/decision tree learning system embedded in an information extraction system.

Learning is also used in the ALT-J/E system for the acquisition of hierarchical semantic knowledge, presented by Yamazaki, Pazzani and Merz. An inductive learning method (FOCL) is first used to learn translation rules whose disjunctions are clustered to form the semantic hierarchy. The frequency of co-occurring terms in a disjunction of rules measures the similarity of terms for the semantic hierarchy. Then the average linkage method for clustering is applied to build the hierarchy. The experimental results showed that learning or updating semantic hierarchies improves the accuracy of learning translation rules.

Moulinier and Ganascia describe the application of an inductive learning system, CHARADE, to the problem of text categorization. Given an attribute-value representation of training examples, this system generates k-DNF rules that cover the positive training examples but not the negative examples. This paper contrasts CHARADE with decision-tree learning algorithms and compares the performance with previously reported results for a decision tree algorithm and a Bayesian classifier on the Reuters categorization corpus. Additional experiments also illustrate the role that redundancy can play in learning effective rule sets.

Inductive logic programming (ILP) is currently a hot topic in machine learning circles, so it is not surprising to see ILP being applied to problems in natural language processing. Zelle and Mooney describe CHILL, an inductive logic programming system, and use it to learn search-control rules for parsing operators. They compare CHILL's performance with that of a naive ILP program that generates a parser without search-control rules. Both systems perform well on a small data set of case-role mapping assignments, but CHILL was much more successful at parsing a larger data set from the Penn Treebank corpus. The paper also discusses how CHILL can be used to generate database queries from sentences.

Mooney and Califf use inductive logic programming to tackle the problem of learning past tenses of English verbs. Their system, FOIDL, induces first-order decision lists that represent rules associated with past tense formation. Key properties are that this system uses only intensional rather than extensional background definitions and does not need explicit negative examples. FOIDL is compared with FOIL [64], IFOIL, and previously reported results.

4.4 Analogical, Rule-based, and Explanation-based Learning

Federici, Pirrelli, and Yvon use analogical-based learning techniques to learn how to pronounce words. The center of their approach is the representation of paradigmatic nodes and links for core components of words that are the same orthographically and phonologically. The paradigmatic nodes are viewed as analogical islands that are generally reliable. This approach is compared with another analogy-based system without paradigms, a decision tree system, and an instance-based learner.

Osborne investigates the role that punctuation might play in learning grammar rules. Osborne's learning system generates grammar rules using a model-based approach to filter out rules that are not consistent with the model and to revise rules so that they become consistent with the model. Since punctuation symbols are often used in natural language to delimit modifiers, separate phrases, and disambiguate sentence structure, Osborne contends that including them in a language model could improve a system's ability to learn an effective grammar.

Joshi and Srinivas describe the application of explanation-based learning (EBL) to speed up the performance of a parser. Their parser is based on an LTAG (Lexicalized Tree Adjoining Grammar) formalism. As input, the learning system uses a parsed corpus that contains dependency and phrase structure information. The learning algorithm is then used to generalize the parses, so that the generalized parses can be used for subsequent sentences. Experimental results show that using the EBL system can substantially reduce the time required for parsing.

A rule-based approach for acquiring compound nouns automatically is described by Jacquemin. He argues that even though there are lexicons available that contain important terms for some domains (e.g., many technical domains), there will always be a need to learn new terms as knowledge evolves. This paper describes a system that begins with a dictionary of terms for a domain and uses a set of rules to infer new terms as they are encountered in text. The rules represent patterns that rewrite the original terms in a different form by recognizing coordinations (e.g., conjunctions formed from the original terms), insertions (new words inserted between the original words), and permutations (e.g., a compound noun transformed into a prepositional phrase). Jacquemin also discusses how conceptual links can be assigned to the new terms and presents empirical results from a medical corpus.

Kaneda and colleagues propose a learning model to support machine translation in the ALT-J/E framework. Verb selection is among the most important problems in machine translation. Their system learns to find appropriate English verbs for Japanese verbs based on the verbs and their semantic case role fillers. Instead of using existing translation examples, they argue that a carefully selected number of hand-made translation rules together with some existing translation examples provides better guidance to the learning model.

Hahn, Klenner, and Schnattinger take a formal approach to concept learning by presenting a terminological representation language. This language is a formalism to support learning the meanings of new words based on predefined knowledge and surrounding context. This approach depends on a large knowledge base of predefined concepts for the domain and associated background knowledge. A key feature of the formalism is that it supports the generation of multiple hypotheses and uses its knowledge sources to sift through and assess competing hypotheses.

4.5 Further Work

To learn more about information extraction techniques and systems, see [47, 36]. Several systems have been developed recently that learn dictionaries for information extraction, such as [43, 67, 74]. Some older systems that incorporated symbolic learning techniques with natural language processing include [1, 29, 9, 37]. Explanation-based learning has also been previously applied to NLP (e.g., see [55]), and rule-based learning techniques have been used to extract information from on-line dictionaries and build knowledge bases automatically [54, 20].

5 Summary and Discussion of General Issues

In this section we step back from the analysis of specific approaches and summarize what we consider to be general issues and trends in the field of learning for natural language processing based on the contributions in this volume and the discussions at the workshop.

5.1 Flat Analysis and Learning

First, there is an important relationship between learning and the underlying representations. In general, the connectionist, statistical, and symbolic approaches described in this volume use flat representations rather than deeply structured representations to support learning. Learning approaches often focus on syntactic/semantic tagging, classification and feature extraction rather than syntactic/semantic analysis using e.g. highly structured HPSG grammars.

5.2 Comparative Evaluation of Different Learning Methods

It is important and necessary to compare different learning approaches according to their strengths and weaknesses. We have seen several examples for such an evaluation of different learning approaches in this volume. It would be interesting to extend such comparisons within the field of learning natural language processing.

5.3 Learning Language Problems from the Real World

We see an important trend emerging for using real-world data for the learning algorithms in language processing. Several years ago, in the communities of natural language processing, connectionism, and machine learning, many smaller "toy" problems or domains were used. However, now that many corpora, lexica, and knowledge bases are available, this opens up many possibilities for further research on learning language in real-world problems and domains.

5.4 Hybrid Approaches for Complex Tasks

In some cases, however, the tasks get so complex that it might not be possible to choose a single best learning technique. This is the case for problems like learning to translate spoken language to another language, where many different modules have to be involved to attack such complex problems. Here we cannot expect that evaluation tests to identify a generally best learning method will succeed. Rather, individual modules can be evaluated and many different learning methods might be useful; sometimes combined with manually-encoded knowledge if it is available. For such complex problems like speech translation or interactive information extraction, the question of a desired hybrid learning architecture is very important and many different connectionist, statistical and symbolic methods may prove useful for solving complex tasks.

References

1. J. R. Anderson. Induction of Augmented Transition Networks. *Cognitive Science*, 1:125–157, 1977.
2. J. A. Barnden and K. J. Holyoak, editors. *Advances in connectionist and neural computation theory*, volume 3. Ablex Publishing Corporation, Norwood, New Jersey, 1994.
3. L. Bookman and R. Sun. Integrating neural and symbolic processes. *Connection Science*, 5:203–204, 1993.
4. H. A. Bourlard and N. Morgan. *Connectionist Speech Recognition.* Kluwer, Boston, 1993.
5. E. Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, 1993.
6. T. Briscoe and J. Caroll. Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. Technical Report TR 224, Computer Laboratory, University of Cambridge, UK, England, June 1991.
7. P.J. Brown, S. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and R. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.
8. J. D. Burger and D. Connolly. Probabilistic resolution of anaphoric reference. In *Probabilistic Approaches to Natural Language*, AAAI Fall Symposium, 1992. AAAI Press.
9. J. G. Carbonell. Towards a self-extending parser. In *Proceedings of the Seventeenth Meeting of the Association for Computational Linguistics*, pages 3–7, 1979.

10. J. G. Carbonell. Derivational analogy: a theory of reconstructive problem solving and expertise acquisition. In *Machine Learning: An Artificial Intelligence Approach.* Morgan Kaufmann, San Mateo, CA, 1986.
11. E. Charniak. *Statistical Language Learning.* MIT Press, Cambridge, MA, 1993.
12. M. H. Christiansen. The (non)necessity of recursion in natural language processing. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 665–670, Indiana University, Bloomington, 1992.
13. K. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), 1990.
14. G. W. Cottrell. A model of lexical access of ambiguous words. In S. I. Small, G. W. Cottrell, and M. K. Tanenhaus, editors, *Lexical Ambiguity Resolution*, pages 179–194. Morgan Kaufmann, San Mateo, CA, 1988.
15. D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140, 1992.
16. G. Cybenko. Approximation by superposition of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.
17. G. DeJong and R. Mooney. Explanation-based learning: an alternative view. *Machine Learning*, 1:145–176, 1986.
18. S. J. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1), 1988.
19. J. Diederich. An explanation component for a connectionist inference system. In *Proceedings of the Ninth European Conference on Artificial Intelligence*, pages 222–227, Stockholm, 1990.
20. W. Dolan, L. Vanderwende, and S. D. Richardson. Automatically deriving structured knowledge bases from on-line dictionaries. In *Proceedings of the First Conference of the Pacific Association for Computational Linguistics*, pages 5–14, 1993.
21. G. Dorffner, editor. *Neural Networks and a New AI.* Chapman and Hall, London, UK, 1995.
22. M. G. Dyer. Symbolic neuroengineering for natural language processing: a multilevel research approach. In J. A. Barnden and J. B. Pollack, editors, *Advances in Connectionist and Neural Computation Theory, Vol.1: High Level Connectionist Models*, pages 32–86. Ablex Publishing Corporation, Norwood, NJ, 1991.
23. J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–221, 1990.
24. J. Feldman. Structured connectionist models and language learning. *Artificial Intelligence Review*, 7(5):301–312, 1993.
25. J. A. Feldman and D. H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.
26. D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.
27. W. N. Francis and H. Kucera. *Manual of Information to Accompany a Standard Corpus of Present-day Edited American English.* Brown University, Department of Linguistics, 1979.
28. R. Garside, G. Leech, and G. Sampson. *The Computational Analysis of English: A Corpus-Based Approach.* Longman, 1983.
29. R. H. Granger. FOUL-UP: A program that figures out meanings of words from context. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 172–178, 1977.
30. G. Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer, Boston, 1994.

31. K. Hammond. CHEF: A model of case-based planning. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 267–271, 1986.
32. J. Henderson. Connectionist syntactic parsing using temporal variable binding. *Journal of Psycholinguistic Research*, 6, 1994.
33. J. A. Hendler. Marker passing over microfeatures: towards a hybrid symbolic connectionist model. *Cognitive Science*, 13:79–106, 1989.
34. K. Hornik, W. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.
35. J. Hughes and E. Atwell. Automatically acquiring a classification of words. In *Proceedings of the IEEE Colloquium on Grammatical Inference*, 1993.
36. P. Jacobs and L. Rau. SCISOR: extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.
37. P. Jacobs and U. Zernik. Acquiring lexical knowledge from text: a case study. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 739–744, 1988.
38. A. N. Jain. Generalization performance in PARSEC - a structured connectionist parsing architecture. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 209–216. Morgan Kaufmann, San Mateo, CA, 1992.
39. D. Jones, editor. *New Methods in Language Processing*. University College London, 1995.
40. M. I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Conference of the Cognitive Science Society*, pages 531–546, Amherst, MA, 1986.
41. M. Kay and M. Röscheisen. Text-translation alignment. *Computational Linguistics*, 18(2), 1993.
42. G. Kempen and T. Vosse. Incremental syntactic tree formation in human sentence processing: a cognitive architecture based on activation decay and simulated annealing. *Connection Science*, 1 (3):273–290, 1989.
43. J. Kim and D. Moldovan. Acquisition of semantic patterns for information extraction from corpora. In *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications*, pages 171–176, 1993.
44. J. Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, (6):225–242, 1992.
45. S. C. Kwasny and K. A. Faisal. Connectionism and determinism in a syntactic parser. In N. Sharkey, editor, *Connectionist Natural Language Processing*, pages 119–162. Lawrence Erlbaum, 1992.
46. G. Leech, R. Garside, and E. Atwell. The automatic grammatical tagging of the LOB corpus. *ICAME News*, 7:13–33, 1983.
47. W. G. Lehnert and B. Sundheim. A performance evaluation of text analysis technologies. *AI Magazine*, 12(3):81–94, 1991.
48. D. M. Magerman. Natural language parsing as statistical pattern recognition. Technical Report PhD thesis, Stanford University, 1994.
49. M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(1), 1993.
50. J. L. McClelland and A. H. Kawamoto. Mechanisms of sentence processing: assigning roles to constituents. In J. L. McClelland and D. E. Rumelhart, editors, *Parallel Distributed Processing*, volume 2, pages 272–326. MIT Press, Cambridge, MA, 1986.

51. J. L. McClelland, D. E. Rumelhart, and G. E. Hinton. The appeal of parallel distributed processing. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 3–44. MIT Press, Cambridge, MA, 1986.
52. R. Miikkulainen. *Subsymbolic Natural Language Processing.* MIT Press, Cambridge, MA, 1993.
53. T. M. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: a unifying view. *Machine Learning*, 1:47–80, 1986.
54. S. Montemagni and L. Vanderwende. Structural patterns vs. string patterns for extracting semantic information from dictionaries. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 546–552, 1992.
55. R. Mooney and G. DeJong. Learning Schemata for Natural Language Processing. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 681–687, 1985.
56. M. Mozer and P. Smolensky. Using relevance to reduce network size automatically. *Connection Science*, 1 (1):3–16, 1989.
57. *Proceedings of the Fourth Message Understanding Conference*, Morgan Kaufmann, San Mateo, CA, 1992.
58. *Proceedings of the Fifth Message Understanding Conference*, Morgan Kaufmann, San Francisco, CA, 1993. Morgan Kaufmann.
59. S. Nirenburg, J. Beale, and I. Domashnev. A full-text experiment in EBMT. In Daniel Jones, editor, *New Methods in Language Processing.* University College London, 1995.
60. J. B. Pollack. On connectionist models of natural language processing. Technical Report PhD thesis, Technical Report MCCS-87-100, New Mexico State University, Las Cruces, NM, 1987.
61. J. B. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46:77–105, 1990.
62. B. W. Porter, R. Bareiss, and R. C. Holte. Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence*, 45, 1990.
63. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:80–106, 1986.
64. J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5(3):239–266, 1990.
65. J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1992.
66. R. G. Reilly and N. E. Sharkey. *Connectionist Approaches to Natural Language Processing.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.
67. E. Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816. AAAI Press/The MIT Press, 1993.
68. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 318–362. MIT Press, Cambridge, MA, 1986.
69. D. E. Rumelhart and J. L. McClelland. PDP models and general issues in cognitive science. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 110–146. MIT Press, Cambridge, MA, 1986.
70. G. Scheler. Learning the semantics of aspect. In D. Jones, editor, *New Methods in Language Processing.* University College London Press, 1995.

71. H. Schütze and Y. Singer. Part-of-speech tagging using a variable context Markov model. In *Proceedings of the Connectionist Models Summer School*, pages 122–129, Boulder. CO, 1993.
72. N. E. Sharkey. A PDP learning approach to natural language understanding. In I. Alexander, editor, *Neural Computing Architectures*, pages 92–116. North Oxford Academic, 1989.
73. F. Smadja. From n-grams to collocations: an evaluation of Xtract. In *Proceedings of 20th Meeting of the Association for Computational Linguistics*, 1991.
74. S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. CRYSTAL: inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1319, 1995.
75. M. F. St. John and J. L. McClelland. Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–257, 1990.
76. R. Sun. Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, pages 241–295, 1995.
77. J. Veronis and N. M. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 1990.
78. D. L. Waltz and J. B. Pollack. Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74, 1985.
79. S. Wermter and V. Weber. Learning fault-tolerant speech parsing with SCREEN. In *Proceedings of the National Conference on Artificial Intelligence*, pages 670–675, Seattle, USA, 1994.
80. S. Wermter. *Hybrid Connectionist Natural Language Processing*. Chapman and Hall, London, UK, 1995.