Learning Sets of Related Concepts: A Shared Task Model

Tim Hume
ICS Dept.
University of California, Irvine
Irvine, CA 92717
hume@interplay.com

Michael J. Pazzani
ICS Dept.
University of California, Irvine
Irvine, CA 92717
pazzani@ics.uci.edu
(714)824-5888
http://www.ics.uci.edu/dir/faculty/AI/pazzani

Abstract

We investigate learning a set of causally related concepts from examples. We show that human subjects make fewer errors and learn more rapidly when the set of concepts is logically consistent. We compare the results of these subjects to subjects learning equivalent concepts that share sets of relevant features, but are not logically consistent. We present a shared-task neural network model simulation of the psychological experimentation.

Introduction

Researchers have investigated how the relevant background knowledge of the learner influences the speed or accuracy of concept learning (e.g., Murphy & Medin 1985, Nakamura 1985, Pazzani 1991, Wattenmaker *et al.* 1986). However, the psychological investigation to date has only explored problems where subjects learn a single concept and the relevant background knowledge is either brought to the experiment by the subject or given in written instructions. In contrast, research in machine learning has addressed issues that occur when learning a set of related concepts. For example, relevant background concepts might be learned inductively from examples before learning concepts that depend upon this knowledge (Pazzani 1990). Here, we report on two experiments in which subjects induce the relevant background knowledge from examples and use this background knowledge to facilitate later learning. The experiments illustrate the importance of learning the relevance of combinations of features, rather than individual features. We model this experiment with shared-task neural networks (Caruana, 1993).

In the first experiment, subjects first induce the relevant background knowledge and then have the opportunity to use this knowledge in later learning. To more closely simulate the real world, we ran a second experiment wherein the subjects induce the relevant background knowledge at the same time as learning the concept that depends on this knowledge. In both experiments, subjects were divided into two groups. One group, the "feature consistency" group, learned a complex concept that shared relevant features with previously learned related concepts, but was not logically consistent with those concepts. Another group, the "logical consistency" group, learned a complex concept that was logically consistent with previously learned related concepts.

Initial Psychological Experimentation

In the first experiment, subjects were asked to imagine that they work for the US Forest Service and were assigned the task of learning to predict years in which there is a severe risk of forest fire danger in the fall. Four concepts had to be learned in the experiment -- one concept in each of four phases. All subjects learned the same 3 background concepts in phases 1-3. Then, for phase 4, they were divided into two groups (the logical consistency group and the feature consistency group) to learn one of two separate concepts which depended on the background concepts.

The first phase of the experiment was designed to minimize the effects of the subjects' domain-specific pre-existing theories by having every subject learn the same concept. In this first phase, subjects had to learn when there is a severe risk of forest fires in the fall given data on rain in the spring and summer. An example of these data is shown in Figure 1. Subjects were given data that indicated that there is a severe risk of forest fires in the fall only when there is both a wet spring and a dry summer. This rule is consistent with the knowledge of most people who live in Southern California. In the remaining phases, when we measure the learning rate and number of errors made by subjects, novel stimuli are used as features to insure that the knowledge was acquired during the experiment.

Next, the subjects were told that the US Forest Service needs to do advance planning, so it cannot wait until the end of summer to predict when there will be a severe risk of fire in the fall. The subjects again examined data from several years. This time, however, the data was from five simulated scientific instruments that are used each January to detect the presence of factors that may be useful in predicting the amount of rain. When one of the instruments detects the presence of a particular factor, it displays a distinctive graph, as shown in Figure 2. Otherwise, a bar is shown to mark the absence of the instrument's graph (see Instrument 3 of Figure 2) Each instrument displays a graph whose shape differs from that of the other instruments. In this second concept learning problem, subjects had to learn to predict from the instrument readings when there would be a rainy spring. All subjects were given data that indicated there would be a wet spring when one particular instrument showed a distinctive graph. All subjects learned a rule of the form "There will be a wet spring when Instrument-A displays a graph," with the instrument corresponding to

Instrument-A selected randomly. This concept will serve as background knowledge for learning the fourth concept.

In the third concept learning problem, subjects learned another piece of background knowledge. Here, subjects had to learn to predict from the instrument readings when there would be a dry summer. All subjects were shown data derived from the rule "There will be a dry summer when Instrument-B or Instrument-C displays a graph."

In the fourth, and final, concept learning problem, subjects had to learn to predict from the instrument readings when there would be a severe risk of fire in the fall. Concepts 1-3 served as background knowledge for this concept. Subjects in the logical consistency group were given data that indicated there would be a severe risk of fire when Instrument-A displayed a graph and when either Instrument-B or Instrument-C (or both) displayed a graph, i.e., $A \wedge (B \vee C)$. This concept is logically consistent with the first three concepts that were learned. Subjects in the feature consistency group were given data that indicated there would be a severe risk of fire when Instrument-C displayed a graph and when either Instrument-B or Instrument-A (or both) displayed a graph, i.e. $C \wedge (B \vee A)$. Although not consistent with the concepts that were learned, this concept shares relevant features with the logical consistency concept.

Subjects. The subjects were 18 male and female undergraduates attending the University of California, Irvine who participated in this experiment to receive extra credit in an introductory psychology course.

Stimuli. The stimuli consisted of data that were displayed on a computer monitor. In the first concept, since there are two two-valued features, 4 distinct stimuli were constructed. In the remaining three concepts, there were 32 distinct stimuli since there are five two-valued features. The stimuli were presented in a random order for each subject.

Procedures. Each subject was shown data on the computer from a single year and asked to make a prediction (e.g., whether there would be a severe risk of fire in the fall) by clicking on a circle next to the word Yes or a circle next to the word No (i.e., using a mouse to move a pointer to the circle and pressing a button on the mouse). Next, the subject clicked on a box labeled Check Answer. While still displaying the data, the computer indicated to the subject whether his answer was the correct answer. If the subject's answer was correct, the subject could click on a box labeled Continue and data from another year was shown. Otherwise, he selected a different answer and clicked on Check Answer again. This process was repeated until the subjects performed at a level that ensured they had learned an accurate approximation to the concept (making no more than one error in any sequence of 24 consecutive trials). The subjects were allowed as much time as they wanted to make their prediction and to view the data after the correct answer was shown. This process of learning a concept to criteria was repeated for each of the four concepts learned. We recorded the number of the last trial on which the subject made an error, the total number of errors made by the subject for each concept, and the number of made on each block of 16 trials. If the subject did not obtain the correct answer after 96 trials, we recorded that the last error was made on trial 96.

Results. Subjects in the logical consistency group required an average of 27.6 trials to learn the fourth concept, while subjects in the feature consistency group required an average of 50.4 trials $t(16) = 1.91$, $p < .05$. Subjects in the logical consistency group made an average of 6.8 errors, while subjects in the feature consistency group made an average of 14.0 errors $t(16) = 2.135$, $p < .05$.

Multiple Concept Learning

In Experiment 1, subjects accurately induced three relevant background concepts, prior to learning a single concept which depended upon those concepts. The order of the concepts is the ideal order for subjects to first acquire knowledge inductively and then use that knowledge in future learning. However, the natural world does not have a benevolent teacher who orders experiences for the learner. To more closely simulate the natural world, in the second experiment, those concepts that had the same stimuli from the first experiment (the last three concepts) are learned at the same time. For each presentation of stimuli, subjects predicted whether there would be a rainy spring, a dry summer, and a severe risk of fire in the fall (see Figure 3). With this exception, Experiment 2 was identical to Experiment 1. For the second learning phase, subjects had to click on all three boxes correctly before proceeding to the next stimuli. We recorded the number of the last trial on which the subject made an error and the total number of errors made by the subject only for the concept that involved predicting whether there would be a severe risk of fire in the fall from the instrument data. In addition, for this concept, we also recorded the number of errors made by the subject on blocks of 16 trials. If the subject did not obtain the correct answer after 128 trials, we recorded that the last error was made on trial 128.

Results. The subjects in the logical consistency group required an average of 77.8 trials to predict whether there would be a severe risk of fire in the fall from the instrument data, while subjects in the feature consistency group required an average of 109.9 trials $t(16) = 1.81$, $p < .05$. In addition, subjects in the logical consistency group made an average of 29.3 errors, while subjects in the feature consistency group made an average of 41.4. errors. This last figure is marginally significant $t(16) = 1.41$, $p < .1$. The results demonstrate that simultaneously learning a set of related concepts is easier when the concepts are logically consistent than when the concepts merely share a set of relevant features. Figure 4a graphs the percentage of errors made by the two groups at predicting a severe risk of fire in the fall from the instrument data as a function of the number of trials. It shows that subjects in logical consistency and feature consistency groups perform similarly until trial 64. After this point, subjects in the logical consistency group make fewer errors than those in the feature consistency group.

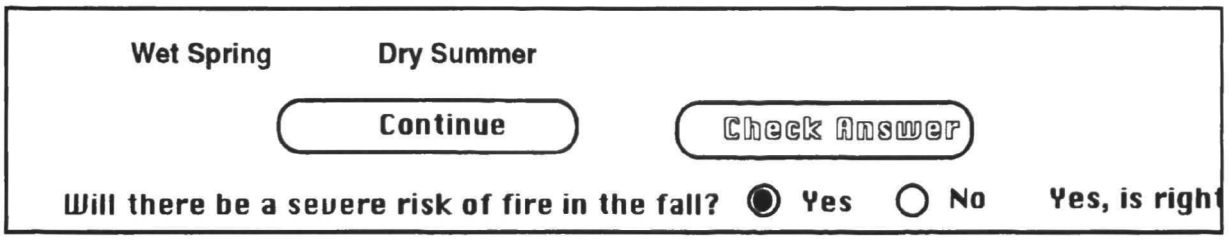

Figure 1. An example of the abstract feature stimuli used for the first concept .

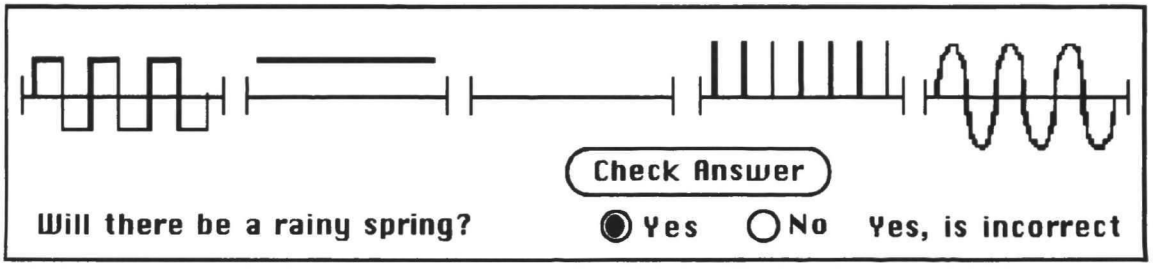

Figure 2. An example of the stimuli used for the second, third and fourth concepts.

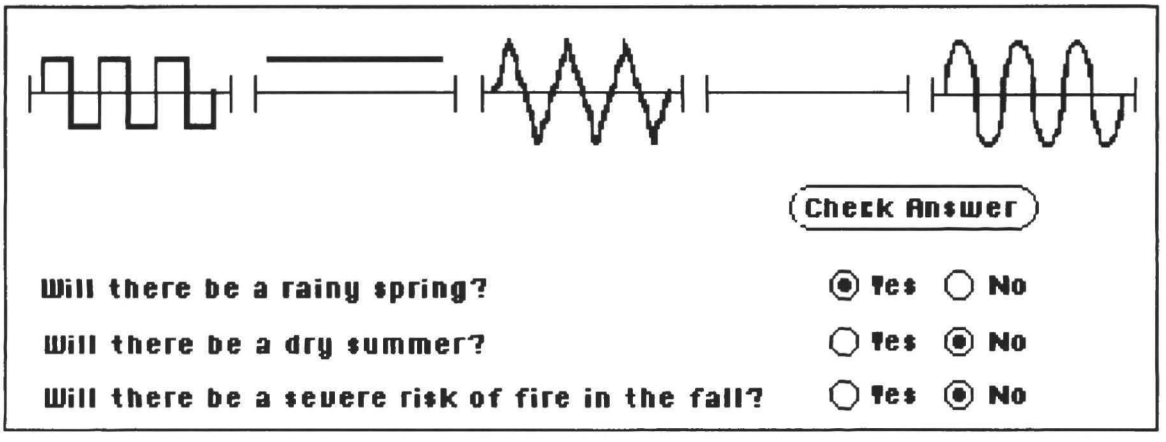

Figure 3. An example of the stimuli used in the second phase of Experiment 2.

Discussion

There are three findings of note in these experiments. First, subjects in the logical consistency condition make fewer errors and require fewer trials to learn. While this finding agrees with our intuition on how people should learn, previous experiments involving background knowledge have not had subjects learn this background knowledge. Furthermore, current cognitive models do not perform in this manner and there is no quantitative data on how background knowledge that is learned inductively influences the learning rate and number of errors made by learners.

Second, learning the relevance of individual features cannot account for these findings. Wisniewski and Medin (1994) use the term selection models to refer to learning models that use prior knowledge to determine which features are relevant. Lien and Cheng (1989) present one such model. Selection models would not be able to explain the results since both the logical consistency and feature consistency groups learn concepts with the same relevant features.

Third, although the subjects in the logical consistency group learn faster and make fewer errors than subjects in the feature consistency group, they learner slower and make more errors than would be predicted by existing computational models of the influence of prior knowledge such as Explanation-based learning (EBL) (Mitchell *et al.* 1986). EBL is a machine learning method that derives concepts from background knowledge. At first, it might seem that EBL would serve as an ideal model of the use of prior knowledge in learning. Its inputs correspond exactly to those items learned in Phases 1-3 of the first experiment, and its output correspond exactly to the concept to be learned in Phase 4. However, there are several problems with EBL as a model of human learning. First, EBL algorithms would learn more quickly than the logical consistency subjects. Since the fourth concept can be deductively derived from the preceding three, EBL would make no errors on this data. Second, EBL cannot function unless the background

knowledge is complete. For example, EBL could not acquire the concepts in Phases 2 and 3 since these are just associations between stimuli and weather predictions.

Modeling with Shared-Task Networks

Here, we propose a model of the psychological experiments using multi-layer neural networks trained with error backpropagation (Rumelhart *et al.* 1986) to learn multiple concepts. First, we would like to make a distinction between sub-task learning and shared-task learning. In sub-task learning, some of the concepts to be learned serve as background concepts for the other concepts to be learned. For instance, in poker, learning the hands *two pair* and *one pair* is a sub-task problem because *one pair* is a background concept for learning *two pair*. Shared-task learning, on the other hand, involves learning concepts that share subordinate concepts. As an example, learning both the hands *two pair* and *full house* require knowing what *one pair* is, but *two pair* and *full house* do not require knowledge of each other.

The network diagrammed on the left side of Figure 5 shows a typical way of using networks to learn sub-task concepts with the network applied to Experiment 1. (Please note that in order to make the diagrams more comprehensible, only some of the connections between nodes are drawn. In an actual network, all the nodes of a hidden layer would be connected to all of its input and output nodes.) The network first learns the section enclosed in the solid line. The two inputs are analogous to the abstract features shown our subjects in the first phase of the experiment. The output is the network's guess at whether or not there will be a severe risk of fire in the fall. Second, the network is trained on the section enclosed in the dashed line. This represents learning the Wet Spring concept. The five inputs (A-E) on the left represent the five instrument displays shown to the human subjects. The output is the network's prediction at whether there will be a wet spring. Third, the Dry Summer concept is trained on the network section enclosed in the dotted line. The same five inputs are used as were used to learn the previous concept. The output is the network's guess at whether there will be a dry summer. The wet spring and dry summer concepts are the sub-tasks the network learns. The final Fire in the Fall concept is represented by training and testing on the entire network. The network uses the five inputs to decide if there will be a severe risk of fire in the fall.

A system such as KBANN (Towell *et al.* 1990) could set up a network like the one on left side of Figure 5, given symbolic inferences rules that represent the knowledge acquired in the first three phases of the experiment. A problem with this method in modeling the experiment is that since the network would already be trained on the three background concepts, it would not require any training to learn the final concept in the logical consistency group of our experiment. This is the same problem that EBL suffers from.

Caruana (1993) has done work on shared-task learning using networks with one hidden layer. The network on the right of Figure 5 is a representation of such a network. A major advantage of this model is that the hidden layer can create new features which can be shared by all of the output units. To model the first experiment, the network first uses the five inputs and only the Wet Spring output unit is trained, i.e., receives feedback on its performance. Second, the same five inputs are used, but only the Dry Summer output unit is trained. Third, the Fire in the Fall output unit is trained and tested using the five inputs.

We performed experiments with shared-task neural nets to see if they could model the results from our psychological experiments since it appeared that this method could learn the combinations of features in addition to feature relevancy. These networks might also be able to combine features and store the combination in the network just as it stores learned knowledge. In both experiments, the first phase used 2 abstract features as stimuli while the later phases used 5 instrument displays. Since the network cannot learn concepts with different forms of inputs, it cannot be trained on the first phase. However, the network can be used to learn the other phases of the experiments. To model the sequential experiment (Experiment 1), the network first uses the 5 inputs and only the Wet Spring output unit is trained, i.e., receives feedback on its performance. Second, the same 5 inputs are used, but only the Dry Summer output unit is trained. Third, the Fire in the Fall output unit is trained and tested using the 5 inputs. Modeling the simultaneous experiment (Experiment 2)is done by training all 3 of the output nodes at the same time, but only using the Fire in the Fall unit for testing.

The logical form of the data was the same as used in the psychological experiments. The first output unit had a value of one when one random feature, say A, had a value of 1. The second output unit had a value of 1 when either (or both) of two other randomly selected features, say B and C, had values of 1. To model the logical consistency group, the third output unit had a value of 1 when feature A had a value of 1 and either feature B or feature C (or both) had a value of 1, i.e. $A \wedge (B \vee C)$. The network used was a feed-forward system with one layer of 20 hidden units. The generalized delta rule was used for training and the logistic function was used for activation. At testing, a network output value greater than 0.5 was treated as a 1 and a value below 0.5 was treated as a 0 to model the forced guessing that was applied to the human subjects. Momentum was set at 0.90 and the learning rate was set at 0.25.

To model Experiment 1, we trained the network to sequentially learn each of the 3 concepts: wet spring, dry summer, and fire in the fall. We first trained the network to learn when an example was a positive example of the wet spring concept, i.e. when the first output unit would have a value of 1 as a function of the 5 features. After each epoch through the training data, the network was tested to see if it could correctly predict the value of the first output unit on at least 31 of the 32 examples. If it could, the network was then trained on learning when the second output unit (dry summer) was true as a function of the 5 features. If it could not reliably predict the first feature, it was trained on another epoch through the data. After it had learned to reliably predict the second output unit, it was trained to predict the third output unit the fire in the fall concept. Data was recorded on how many epochs the network took to learn the

final concept. The process of learning each concept sequentially was repeated 50 times.

The network required an average of 5.96 epochs, or 190.72 trials, to learn the logical consistency set, while it took significantly longer, 8.50 epochs or 272.00 trials, to learn the feature consistency set, $t(98) = 6.06$, $p < .05$. Similar to the human subjects, this network sequentially learned the set of concepts more easily when it was logically consistent than when the concepts merely share features.

To model Experiment 2, we trained the network to simultaneously learn all three concepts. The network was trained on all 3 of the concepts, but was tested only on the third concept. After each epoch through the training data, the network was tested to see if it could correctly predict the value of the third feature on at least 31 of the 32 examples. If it could, then training stopped; otherwise, it was trained for another epoch. Data was kept on how many errors the network made on each epoch and on which epoch the network learned the final concept. The process of learning the concepts was repeated 50 times.

The neural net required an average of 7.12 epochs, or 227.84 trials, to learn the logical consistency set, while it took significantly longer, 9.66 epochs or 309.12 trials, to learn the feature consistency set, $t(98) = 5.039$, $p < .05$. Similar to the human subjects, this network simultaneously learned the set of concepts more easily when it was logically consistent than when the concepts merely share features. Figure 4b graphs the percentage of errors made on the two sets as a function of the number of epochs. It shows that after the second epoch, the graph is similar to Figure 4a. On the logically consistent condition, the network becomes accurate with fewer training epochs.

Shared task networks are able to model these results because they can create new abstract features and use these features to influence learning other concepts. The network requires some training to determine how to use these abstract features, but less training than would be required if new concepts were not consistent with the concepts learned earlier. The shared task network is an example of what Wisniewski and Medin (1994) call a tightly coupled model. Prior knowledge, in this case created by prior learning, selects the relevance of features (by having higher weights on some connections), and creates new features (as represented in the hidden units). Furthermore, feedback during learning one concept can change the features or strengths of the hidden units used by other concepts.

Conclusions

Although the general topic of learning a series of concepts has been discussed, previous research has focused on attentional phenomena such as the intradimensional and extradimensional shift in which subsequent concepts share related features with prior concepts. However, these approaches consider sets of arbitrary groups of concepts rather than concepts that are causally related. Waldmann and Holyoak (1990) argue that the causal induction process differs from the learning process used to acquire arbitrary concepts. In particular, we show that concepts acquired by induction in one phase of an experiment influence later learning in much the same manner as concepts acquired by reading written instructions or prior background concepts.

We have focused on how prior knowledge facilitates learning. We should also point out that incorrect prior knowledge may also hinder learning by providing misconceptions (Chi, Slotta & de Leeuw, 1994). It is only when prior knowledge is compatible with the new knowledge to be acquired that we anticipate a positive effect.

Classical concepts that consistent of sets of necessary and sufficient features have several flaws. Few concepts people encounter have such rigorous logical definitions (Rosch, 1978). More recently, it has become apparent that concepts do not exist and are not learned in isolation. Here, we have presented quantitative results on how induced background knowledge influence the rate of learning and the number of errors made during learning. While we have found that having relevant, correct background knowledge facilitates learning, it does not eliminate the need for learning. That is, unlike previous learning models, when subjects have learned rules corresponding to "$A \rightarrow$ WetSpring," "$B \vee C \rightarrow$ DrySummmer" and "WetSpring $\wedge$ DrySummmer $\rightarrow$ FireInFall" they do not automatically know that "$A \wedge (B \vee C) \rightarrow$ FireInFall." We believe that one flaw in previous learning models that use prior knowledge is that the equate an explanation with a logical proof, and use rules that have necessary and sufficient preconditions. Such rules may be as rare in the real world and as cognitively implausible as concepts that consistent of necessary and sufficient definitions.

Acknowledgments

The research reported here was supported in part by NSF grant IRI-9310413. We thank Kamal Ali, Dorrit Billman, Cliff Brunk, Piew Datta, Dennis Kibler, Chris Merz, Brian Ross, and David Schulenburg for comments on various phases of this work.

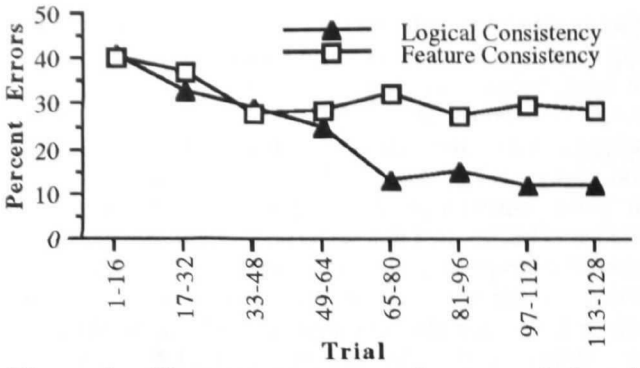

Figure 4a. The mean percentage of errors made by subjects in the logical consistency and feature consistency groups as a function of the trial in Experiment 2.

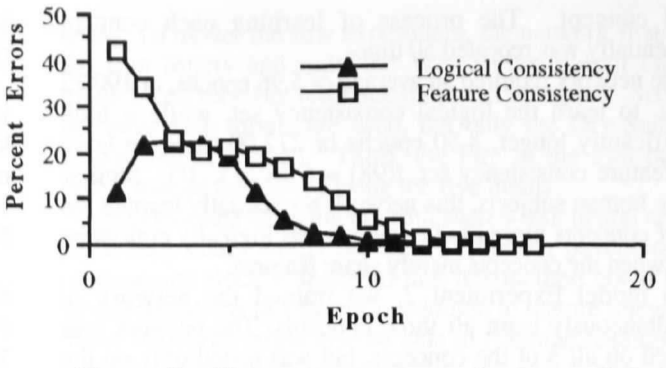

Figure 4b. The mean percentage of errors made by the neural network in the logical consistency and feature consistency groups as a function of the epoch.

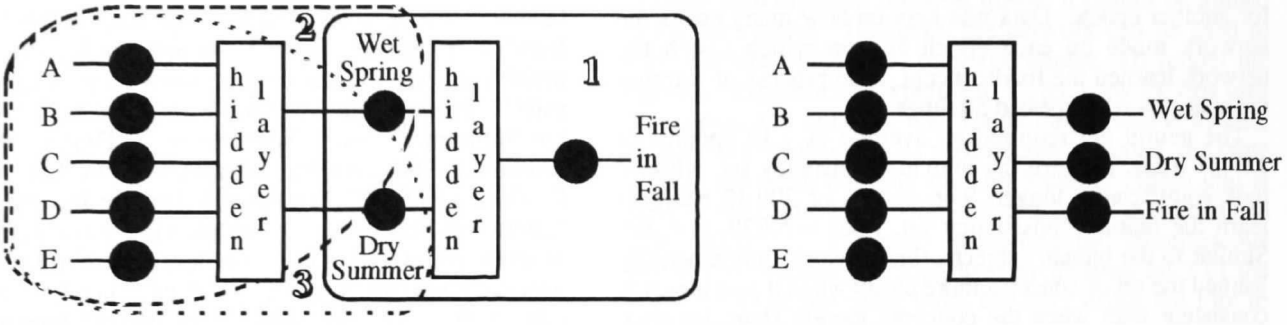

Figure 5. Neural network diagrams. The network on the left is a sub-task learning model. The network on the right is a shared-task learning model.

References

Caruana, R. (1992). Multitask learning: A knowledge-based source of inductive bias. *Proceedings of the Tenth International Machine Learning Conference* (pp. 41-48). San Mateo, CA: Morgan Kaufman.

Chi, M., Slotta, J. & de Leeuw, N. (1994). From theories to processes: A theory of conceptual changes for learning science concepts. *Learning and Instruction, 4,* 27-43.

Lien, Y., & Cheng, P. (1989). A framework for psychological induction: Integrating the power law and covariation views. *The Eleventh Annual Conference of the Cognitive Science Society* (pp. 729-733). Ann Arbor, MI: Lawrence Erlbaum Associates, Inc.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based learning: A unifying view. *Machine Learning, Vol. 1(1).*

Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 3.

Nakamura, G. (1985). Knowledge-based classification of ill-defined categories. *Memory & Cognition. 13,* 377-84.

Pazzani, M. (1990) *Creating a memory of causal relationships: An integration of empirical and explanation-based learning methods.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pazzani, M. (1991). The influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory & Cognition, 17, 3,* 416-32.

Rosch E. (1978). Principles of categorization. In *Cognition and categorization* (Ed.), Rosch E. & Lloyd B.. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by backpropagating errors. In: Rumelhart, D., McClelland, J. (eds.), *Parallel Distributed Processing*, Cambridge, MA: MIT Press.

Towell, G. Shavlik, J. & Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based neural networks. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 861-66). Cambridge, MA: MIT Press.

Waldmann, M. & Holyoak, K. (1990). Can causal induction be reduced to associative learning? *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* Cambridge, MA: Lawrence Erlbaum.

Wattenmaker, W., Dewey, G., Murphy, T., & Medin, D. (1986). Linear separability and concept learning: Context, relational properties and concept naturalness. Cognitive Psychology, 18, 158-194.

Wisniewski, E. & Medin, D. (1994). On the interaction of data and theory in concept learning. Cognitive Science, 18, 221-282.