Working Notes of the 1992 AAAI Workshop on Constraining Learning with Prior Knowledge, San Jose, CA. Also in Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, pages 743–748.

Using Cognitive Biases to Guide Feature Set Selection

Claire Cardie
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
E-mail: cardie@cs.umass.edu

Abstract

Although learning is a cognitive task, machine learning algorithms, in general, fail to take advantage of existing psychological limitations. In this paper, we examine three well-known cognitive biases: 1) the tendency to rely on the most recent information, 2) the heightened accessibility of the subject of a sentence, and 3) short term memory limitations. In a series of experiments, we modify a baseline instance representation in response to these limitations and show that the overall performance of the learning algorithm improves as increasingly more cognitive biases and limitations are incorporated into the representation.

1 Introduction

Inductive concept acquisition has always been of primary interest for researchers in the field of machine learning (see [Michalski, Carbonell, & Mitchell, 1983],[Michalski, Carbonell, & Mitchell, 1986],[Michalski, & Kodratoff, 1990]). In this task, a system typically learns one or more concepts by analyzing a set of examples (and possibly counterexamples) of the concepts. In fact, a number of systems for the acquisition of concepts now exist (e.g., ID3 [Quinlan, 1979], ARCH [Winston, 1975], COBWEB [Fisher, 1987], UNIMEM [Lebowitz, 1987]). Independently, psychologists, psycholinguists, and cognitive scientists have examined the effects of numerous psychological limitations on human information processing. However, despite the fact that concept learning is a basic cognitive task, most machine learning systems for concept formation fail to exploit these limitations and make no attempt to model human concept learning.

In this paper, we show that the explicit encoding of known cognitive biases into the training instance representation can improve the performance of the learning algorithm for cognitively- based learning tasks. More specifically, we use a well-known concept acquisition system and focus on a single learning task from the field of natural language processing (NLP). After training the system using a baseline instance representation, we modify the representation in response to three cognitive biases: 1) the tendency to rely on the most recent information, 2) the heightened accessibility of the subject of a sentence, and 3) short term memory limitations. In a series of experiments, we compare each of the modified instance representations to the baseline.

2 Finding the Antecedents of Relative Pronouns

Our task for the machine learning system is the following: Given a sentence with the relative pronoun "who," learn to recognize the phrase or phrases that represent the relative pronoun's antecedent. Finding the antecedents of relative pronouns is a crucial task for natural language systems because the antecedent must be carried to the subsequent clause where it implicitly fills the actor role.[1] Consider, for example, the following:

Igor shook hands with *the skater* **who** beat him in the race.

A correct semantic interpretation of this sentence should include the fact that "the skater" is the actor of "beat" even though the phrase does not appear in the embedded clause. Only after the natural language system associates "the skater" with "who" can it carry the constituent across the clause boundary to make this inference. Locating the antecedent of "who" may initially appear to be an easy problem because the antecedent often immediately precedes the word "who." Unfortunately, this is not always the case, as shown in S1 and S2 of Figure 1. Even when the antecedent does immediately precede the relative pronoun, it does not appear in a consistent syntactic constituent. In S3, for example, the antecedent is the subject of the preceding clause; in S4, it is the direct object; in S5, it is the object of a preposition. Furthermore, the antecedent of "who" may contain more than one phrase. In S6, for example, the antecedent is a conjunction of three phrases and in S7, either "our sponsors" or its appositive "Gatorade and GE" is a semantically valid antecedent. Occasionally, there is no apparent antecedent at all (e.g., S8).

S1. *The woman* from Philadelphia who played soccer was my sister.
S2. I spoke to *the man* in the black shirt and green hat over in the far corner of the room who demanded to meet the skiers.
S3. *The skater* who won the medal was from Japan.
S4. I saw *the skater* who won the medal.
S5. Igor ate dinner with *the skater* who won the medal.
S6. I'd like to thank *Nike, Reebok, and Adidas,* who provided the uniforms.
S7. I'd like to thank *our sponsors, Gatorade and GE,* who provide financial support.
S8. We wondered who would win the race.

Figure 1: **Antecedents of "who"**

Despite these ambiguities, we will describe how a machine learning system can learn to locate the antecedent of "who" given a description of the clause that precedes it. In effect, we are teaching the system to recognize the "relative pronoun antecedent" concept. More importantly, we will show that performance of the learning system improves as the instance description includes increasingly more cognitive limitations and cognitive biases.

3 COBWEB and the Representation of Training Instances

For our experiments we chose COBWEB [Fisher, 1987] › a well-known concept formation system that is one of a relatively small number of concept acquisition systems designed to model some aspects of human concept learning.[2] Given a set of training instances, COBWEB discovers a classification scheme that covers the instances. Instead of forming concepts at a single level of abstraction, however, COBWEB organizes instances into a classification hierarchy where leaves represent instances and internal nodes represent concepts that increase in generality as they approach the root of the tree. In addition, COBWEBÖs construction of the hierarchy is cognitively economical in that new objects are incrementally added to the hierarchy as they arrive. To evaluate the concepts it creates, COBWEB employs the *category utility* metric [Gluck, & Corter, 1985] › a measure developed in psychological studies of basic level categories. As a result, the hierarchies constructed by COBWEB account for prototypicality effects and basic level phenomena

[1] In practice, the antecedent of "who" sometimes fills semantic roles other than the actor.
[2] The COBWEB3 system was provided by the kind folks at NASA, Ames.

observed in humans ([Jolicoeur, Gluck, & Kosslyn, 1984], [Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976], [Rosch, et al., 1976]).

COBWEB takes as input a set of training instances described as a list of attribute-value pairs. Because the antecedent of a relative pronoun usually appears as one or more phrases in the clause preceding "who," the attribute-value pairs in each training case represent the constituents that precede "who." At first glance, it may seem that only syntactic information needs to be encoded. However, finding the antecedent of a relative pronoun actually requires the assimilation of syntactic and semantic knowledge. For this reason, each *constituent attribute-value pair* takes the following form:

- The attribute describes the syntactic class and position of the phrase.
- The value provides its semantic classification.

Consider, for example, the sentences in Figure 2. In the training instance for S1, we represent "the man" with the attribute-value pair *(s human)* because it is the subject of the sentence and the noun "man" is human. We represent "from Oklahoma" with the pair *(s-pp1 location)* because it is the first prepositional phrase that follows the subject and "Oklahoma" is a location. All noun phrases are described by one of seven general semantic features: human, proper-name, location, entity, physical-target, organization, and weapon.[3] When clauses contain conjunctions and appositives, each phrase is labelled separately. In S2, for example, the real direct object of "thank" is the conjunction "Nike and Reebok." However, in our instance representation, "Nike" is tagged as the direct object *(do)* and "Reebok" as the first noun phrase that follows the direct object *(do-np1)*.[4] For verb phrases, we currently note only the presence or absence of a verb using the values *t* and *nil*, respectively.

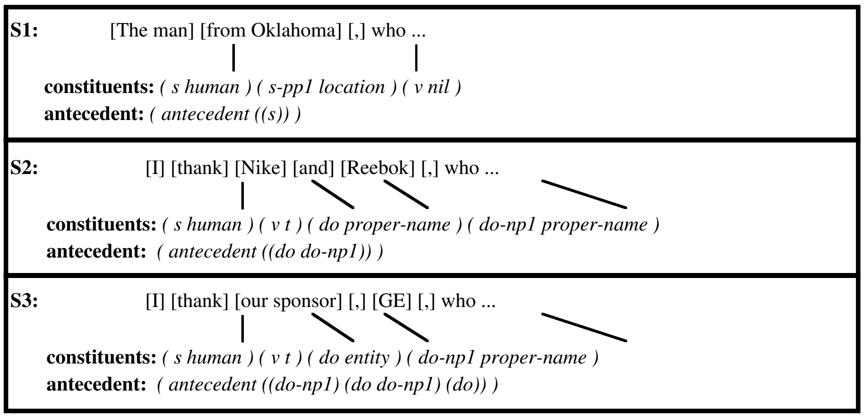


Figure 2: **Training Cases**

In addition to the constituent attribute-value pairs in every training instance, we include information about the correct antecedent in the form of an *antecedent attribute-value pair*.[5] The value of the antecedent attribute is a list of the constituent attributes that represent the location of the antecedent or *(none)* if no antecedent can be found. In S1, for example, the antecedent of "who" is "the man." Because this phrase is represented as the constituent pair *(s human)*, the value of the antecedent attribute is *(s)*. Sometimes, however, the antecedent is actually a conjunction of constituents. In these cases, we represent the antecedent as a list

[3]These features are specific to the domain from which the training instances were extracted.

[4]In a separate paper that focuses on this learning task from the NLP perspective, we explain this representational decision in detail. In general, NLP systems do not reliably handle complex conjunctions and appositives. They can, however, accurately locate lower level phrases like individual noun phrases, verbs, and prepositional phrases. As a result, we let the the machine learning system recognize conjunctions and appositives and allow the NLP system that generates the training instances to ignore these tasks.

[5]COBWEB is often used to perform unsupervised learning. However, many applications of COBWEB, including our own, encode pseudo-supervisory information as part of the instance representation.

of the constituent attributes associated with each element of the conjunction. Look, for example, at sentence S2. Because “who" refers to the conjunction “Nike and Reebok," the antecedent is described as *(do do-np1)*. S3 shows yet another variation of the antecedent attribute-value pair. In this example, an appositive creates three semantically equivalent antecedent values, all of which become part of the antecedent feature: 1) “GE" — *(do-np1)* , 2) “our sponsor" — *(do)*, and 3) “our sponsor, GE" — *(do do-np1)*.

After training, we use the resulting hierarchy to predict the antecedent of “who" in new contexts. Given a new instance to classify, COBWEB retrieves the most specific concept that adequately describes the instance. Then, the antecedent of the retrieved concept guides selection of the antecedent for the novel case. Given the test instance in Figure 3, for example, COBWEB retrieves an instance that specifies *do* as the location of the antecedent. Therefore, we choose the contents of the *do* constituent — “the hardliners" — as the antecedent in the novel case. Sometimes, however, COBWEB retrieves a concept that lists more than one option as the antecedent. In these cases, we choose the option that appeared most often in the underlying instance(s) and whose constituents overlap with those in the current context. (For a description of better, but more complicated heuristics, see [Cardie, 1992].)

Novel instance: [It] [was] [the hardliners] [in Congress] who...

(s entity) (v t) (do human) (do-pp1 entity)

Antecedent of Retrieved Instance: ((do))
Antecedent of Novel Instance: (do) = "the hardliners"

Figure 3: **Instance Retrieval**

4 The Baseline Experiments

We tested this baseline instance representation by extracting all examples of “who" from 3 sets of 50 texts from the MUC-3 corpus.[6] In each of 3 experiments, 2 sets were used for training and the third reserved for testing. The results are shown in Figure 4 and indicate that, using the baseline instance representation, COBWEB can find the correct antecedent of “who" an average of 59% of the time. In the next two sections, we modify this baseline instance representation in response to three cognitive biases and show the results of these modifications on COBWEBÖs performance.

Exp #	Training Sets (# instances)	Test Set (# instances)	Baseline Rep
1	set1 + set2 (170)	set3 (71)	63%
2	set2 + set3 (159)	set1 (82)	47%
3	set1 + set3 (153)	set2 (88)	66%

Figure 4: **Baseline Results (% correct)**

[6]The MUC-3 corpus consists of 1500 texts (e.g., newspaper articles, TV news reports, radio broadcasts) containing information about Latin American terrorism and was developed for use in the Third Message Understanding System Evaluation and Message Understanding Conference [Sundheim, May,1991].

5 Incorporating the Recency Bias

In processing language, people consistently show a bias towards the use of the most recent information (e.g., [Kimball, 1973], [Frazier, & Fodor, 1978], [Gibson, 1990]). In particular, the mechanisms people use for finding the antecedents of pronouns and missing subjects have been investigated in a series of recent experiments (see [Nicol, 1988]). The results show that in locating antecedents during language processing, people consider all noun phrases preceding the pronoun starting with the most recent noun phrase and working backwards to the most distant noun phrase.

We translate this recency bias into representational changes for the training instances in two ways. First, we label the constituent attribute-value pairs with respect to the relative pronoun. This establishes a right-to-left labelling rather than the left-to-right labelling of the baseline. In Figure 55, for example, "in Congress" receives the attribute *pp1* because it is a prepositional phrase one position to the left of "who." Similarly, "the hardliners" receives the attribute *np2* because it is a noun phrase two positions to the left of "who." Notice, however, that the subject of the sentence retains its original *s* attribute. We based this decision on studies that indicate that the subject of a sentence remains highly accessible even at the end of a sentence (e.g., [Gernsbacher, Hargreaves, & Beeman, 1989]). In addition, this new right-to-left labelling tags the antecedents in both "it was *the hardliners* in Congress, who..." and "I heard from *the hardliners* in Congress, who..." with the same attribute (i.e., *np2*). In the baseline representation, the antecedents in each example retain distinct attributes › *do* and *v-pp1*, respectively.

Sentence: [It] [was] [the hardliners] [in Congress] who...

Baseline Representation: *(s entity) (v t) (do human) (do-pp1 entity) (antecedent ((do)))*
Right-to-Left Labelling: *(s entity) (v t) (np2 human) (pp1 entity) (antecedent ((np2)))*
Duplicate Information: *(s entity) (v t) (do human) (do-pp1 entity) (most-recent entity) (part-of-speech prep-phrase) (antecedent ((do))*

Figure 5: **Incorporating the Recency Bias**

Alternatively, given the baseline instance representation, we can incorporate the recency bias by including more than one attribute-value pair for the most recent information. Figure 5 also shows this second representational change. The most recent constituent,"in Congress," is represented three times: 1) as a constituent attribute-value pair — *(do-pp1 entity)*, 2) as the most recent constituent — *(most-recent entity)*, and 3) via its part of speech — *(part-of-speech prep-phrase)*. In this representation, we also allow the antecedent attribute-value pair to refer to the more general *most-recent* constituent rather than the equivalent, but more specific, constituent attribute-value pair. If, for example, the antecedent in Figure 5 had been do-pp1, it would become most-recent in the new representation.

Exp #	Training Sets	Test Set	Baseline	MR1: R-to-L	MR2: Redundant	MR1+MR2
1	set1+set2 (170)	set3 (71)	63%	75%	83%	84%
2	set2+set3 (159)	set1 (82)	47%	62%	65%	73%
3	set1+set3 (153)	set2 (88)	66%	66%	71%	74%

Figure 6: **Experiments Using the Recency Bias (% correct)**

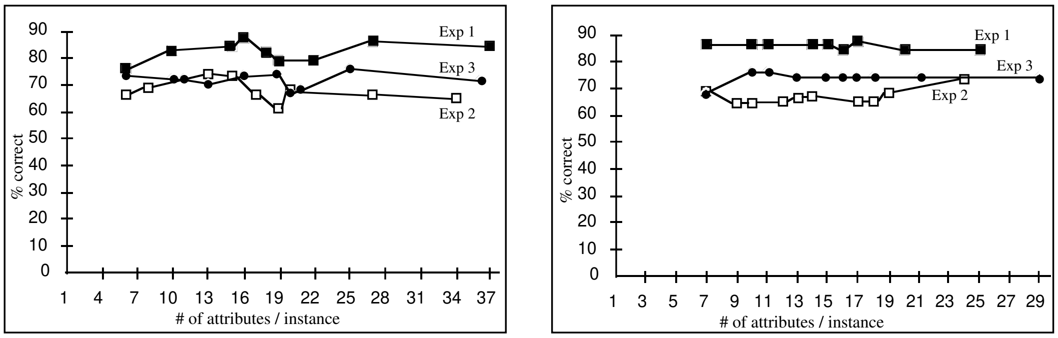
The results of experiments that use each of these representations separately and in a combined form are shown in Figure 6. In this table, the MR1 representation used the right-to-left labelling, the MR2 representation included extra information about the most recent constituent, and the MR1+MR2 representation combined both the right-to-left labelling and the duplicate information format. In general, it is clear that incorporating the recency bias into the instance representation improves performance. On average, the right-to-left labelling increased the percentage of correctly identified antecedents from 59% to 68% while duplication of recent information increased the percentage correct to 73%. The best results, however, occurred using the combined representation, where the percentage correct increased to an average of 77%.

6 Incorporating the Short Term Memory Bias

Psychological studies have determined that people can remember at most seven plus or minus two facts at any one time ([Miller, 1956]). More recently, Daneman and Carpenter ([Daneman, & Carpenter, 1980], [Daneman, & Carpenter, 1983]) show that working memory capacity affects a subject's ability to find the referents of pronouns over varying distances. Also, King and Just [King, & Just, 1991] show that differences in working memory capacity can cause differences in the reading time and the comprehension of certain classes of relative clauses. COBWEB, however, clearly does not make use of short term memory limitations in that each training and test instance has to be normalized with respect to all attributes across the training instances. For the baseline representation, this normalization resulted in instances of 35 attribute-value pairs as compared to an average of 5 attribute-value pairs in the original, unnormalized instances.

In an attempt to incorporate short term memory (STM) limitations, we ran a series of experiments using instances with successively fewer features. We let n = 1,2,3,4,5,7,9,15,20, and 50, and included in the training and test instances only those attributes that occurred at least n times in the training set. When this STM cutoff was applied to the baseline representation, the performance of the learning algorithm gradually declined as n increased. The percentage correct declined from 63% to 37%, from 47% to 19%, and from 66% to 41% for experiments 1, 2, and 3, respectively. Although the STM cutoff did not improve performance for the MR1 training sets, the decline in percentage correct was not nearly so drastic. For experiment 1 (originally 75% hit rate), the percentage correct never dropped below 69%. For experiment 2, results ranged from 62% (with no cutoff) to 49%; and in experiment 3 (originally 66% correct), results ranged from 51% to 67%.

The results of the STM cutoff for the instance representations of MR2 and MR1+2 are shown in Figure 7. In these experiments, the STM bias actually improved COBWEBÖs performance. In the MR2 experiments, the original hit rate for experiment 1 increased from 83instance) to 87from 65instance). In experiment 3, the percentage correct increased from 71the original representation to 76the MR1+2 instance representation. There were increases from 8487instance) for experiments 1 and 3, respectively. Performance for experiment 2, however, declined.



(a) Applying STM Bias to MR2

(a) Applying STM Bias to MR1+2

Figure 7: **STM Bias Results**

7 Conclusions

There is evidence from recent studies in psycholinguistics that some language learning tasks are not distinct from the general cognitive task of concept learning [Mc Donald, & MacWhinney, 1991]. Our machine learning results for the relative pronoun antecedent task appear to corroborate these studies. It is clear that further experimentation is required to explore the effects of additional cognitive limitations, to determine the biases that work well together, and to find the correct parameters for those biases. However, based on the experiments presented in sections 4 through 6, we conclude that explicit incorporation of cognitive biases into the instance representation can greatly improve learning algorithm performance at least for some cognitvely-based tasks.

8 References

Cardie, C. (1992).Learning to Disambiguate Relative Pronouns. Proceedings, Ninth National Conference on Artificial Intelligence. San Jose, CA.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19,450-466.

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9, 561-584.

Fisher, D. H. (1987). Knowledge Acquisition Via Incremental Conceptual Clustering. Machine Learning, 2, 139-172.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. Cognition, 6, 291-325.

Gernsbacher, M. A., Hargreaves, D. J., & Beeman, M. (1989). Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. Journal of Memory and Language, 28, 735-755.

Gibson, E. (1990).Recency preferences and garden-path effects. Proceedings, Twelfth Annual Conference of the Cognitive Science Society. Cambridge, MA.

Gluck, M. A., & Corter, J. E. (1985).Information, uncertainty, and the utility of categories. Proceedings, Seventh Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates.

Jolicoeur, P., Gluck, M., & Kosslyn, S. (1984). Pictures and Names: Making the Connection. Cognitive Psychology, 16, 243-275.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. Cognition, 2, 15-47.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. Journal of Memory and Language, 30, 580-602.

Lebowitz, M. (1987). Experiments with Incremental Concept Formation: UNIMEM. Machine Learning, 2, 103-138.

Mc Donald, J., & MacWhinney, B. (1991). Levels of learning: a comparison of concept formation and language acquisition. Journal of Memory and Language, 30(4), 407-430.

Michalski, R., & Kodratoff, Y. (1990). Machine Learning: An Artificial Intelligence Approach. San Mateo: Morgan Kaufmann,

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). Machine Learning: An Artificial Intelligence

Approach. Palo Alto: Morgan Kaufmann,

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1986). Machine Learning: An Artificial Intelligence Approach. Los Altos: Morgan Kaufmann,

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psycol. Review, 63(1),

Nicol, J. (1988). Coreference processing during sentence comprehension. MIT.

Quinlan, J. R. (1979). Discovering Rules from Large Collections of Examples: A Case Study. In D. Michie (Ed.), Expert Systems in the Microelectronics Age Edinburgh: Edinburgh University Press.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic Objects in Natural Categories. Cognitive Psychology, 8, 382-439.

Winston, P. H. (1975). Learning Structural Descriptions from Examples. In P. H. Winston (Ed.), The Psychology of Computer Vision New York: Mc Graw-Hill.