MODELING AND DIAGNOSING DOMAIN KNOWLEDGE

USING LATENT SEMANTIC INDEXING

Jared T. Freeman
Aptima, Inc., Woburn, MA
Bryan T. Thompson, and Marvin S. Cohen
Cognitive Technologies, Inc., Arlington, VA

Contact: freeman@aptima.com

Submitted to:
Interactive Learning Environments
Special Issue on Latent Semantic Indexing
Editor: Joe Psotka

ABSTRACT

A Latent Semantic Index (LSI) was constructed from arguments made by Navy officers concerning events in an Anti-Air Warfare scenario. A model based on LSI factor values predicted level of domain expertise with 89% accuracy. The LSI factor space was reduced using MDS to five dimensions: aircraft route, aircraft response, kinematics, localization, and an unclassifiable element. Arguments in the localization category were reliably more common among officers with the greatest expertise. Automated classification of arguments into these elements achieved 84% accuracy. LSI may be a useful tool for automating aspects of modeling expertise and diagnosing knowledge deficiencies.

Keywords: LSI, cognitive model, assessment, diagnosis
INTRODUCTION

Situation Understanding

Superior decision making is driven, in part, by accurate knowledge of events, or situation awareness (Endsley, 1997), and by appropriate interpretation of events, or situation assessment (Cohen, Freeman, and Thompson, 1998; Hastie and Pennington, 1991). In some domains, there is an exceptionally high value placed on these products of skilled cognition, which we shall call situation understanding. One such domain is naval anti-air warfare (AAW). Two disastrous events illustrate the importance of accurate situation understanding. In 1988, the U.S.S. Vincennes mistook an oncoming commercial airliner for a hostile fighter aircraft and destroyed it, killing 290 civilians. In 1987, the U.S.S. Stark failed to act against an Iraqi fighter aircraft before it fired two missiles at the ship. Thirty-seven sailors died.

One way to minimize the frequency of calamities such as this, whether in AAW or other domains, is to improve training. Specifically, we need to model situation understanding among experts and non-experts, and diagnose situation understanding among trainees during instruction and practice. However, cognitive modeling is a demanding and costly process. Interview methods such as the critical incident technique elicit data that are often valuable, but analysis of these data is slow and costly. Laboratory techniques such as card sorting are more rapid, but they present domain practitioners with tasks that have little or no relevance to the domain. They lack face validity, at a minimum. Diagnosing complex skills is also a notoriously difficult problem, one that has been successfully tackled mainly in well-structured domains such as mathematics (Jones and van Lehn, 1994; Shute, 1992).

In the work reported here, we apply Latent Semantic Indexing (LSI) to automate diagnosis and modeling of situation understanding in AAW. We analyze a corpus of arguments by U.S. Navy officers concerning the intent of suspect aircraft in a high-fidelity tactical situation. LSI, in combination with other statistical tools, produces relatively accurate assessments of domain expertise, provides useful insights into mental models of the domain, and can be used to diagnose specific deficiencies in situation understanding.

Arguments as a Window onto Situation Understanding

A fundamental challenge in modeling and assessing performance in complex domains is finding a window onto cognition. Particularly in team settings, such as AAW, one such window is argument; that is, statements of evidence, conclusions, rebuttals, and counterarguments. Toulmin (Toulmin, Rieke, and Janik, 1984) proposed that argumentation is a fundamental form of discourse in the range of human enterprises he studied: business, law, management, the arts, and ethics. Not surprisingly, one also finds evidence of argument in the transcripts of AAW incidents and in interviews with AAW officers. These often take the form of discussions between senior officers concerning alternative explanations of ambiguous events, such as those that confronted the commanders of the Stark and Vincennes.
Such arguments are not an epiphenomenal aspect of decision making. Green (1994) asked subjects to explain their choices in Wason's (1968) four card selection task and found that those who answered correctly were the most likely to indicate that their actions tested the truth or falsity of the given rule (rather than that they flipped the cards that were mentioned in the rule, for example). This result was inconsistent with the notion that the individuals were merely justifying their actions post hoc, and it was consistent with the thesis that individuals generate and test arguments during the decision-making process. Green (1996) offered further support for this claim in a study of reasoning about a case of food poisoning, in which he found that the arguments subjects offered predicted the fines they levied against a restaurant, and that presenting subjects with an argument biased their decisions except when they explicitly rebutted that argument.

Shafir, Simonson, and Tversky (1993) predicted that the absence of evidence and a warrant justifying immediate action would lead individuals to defer a decision. They asked three groups of individuals to decide whether to buy a vacation package at a low price, not buy it, or reserve it with a small down-payment. One group was told they had just passed an exam, another learned they had failed it, and the third was told that their results would be known in two days. While half of the subjects who knew their results purchased the package immediately, two thirds of those awaiting their exam results deferred the decision. The researchers suggested that subjects relied on the presence of evidence (exam results) and a warrant (good performance justifies a reward; bad performance deserves compensation) to justify the decision to purchase the vacation immediately. Subsequently, Green (1996) asked subjects to justify their choices in the same decision task and confirmed the supposition of Shafir, et al., that subjects used the presence or absence of a justification for vacationing to drive their purchase decision.

Experimentally manipulating the skills of argument (rather than the availability of evidence) also influences decision making. Cohen, et al. (1998) confirmed that training Navy officers to construct and critique an argument regarding the assessment of an air radar track improved the accuracy of decisions and the processes of decision making. Freeman, et al. (1997), demonstrated that short and focused training concerning the use of arguments plus technology to help diagram arguments improved the tactical decisions of military officers. While this was only a pilot study (n = 11), strong trends indicated that the training improved tactical decisions, strengthened the arguments offered in defense of assessments, and increased consideration of supporting evidence, conflicting evidence, assumptions, and gaps in the evidence. Training in argument

---

1 In a typical version of the Wason card selection task, subjects are shown four cards with a letter on one side and a number on the other (A, K, 4, and 7). They are asked to determine in as few flips of the cards as possible whether the following claim is true: "If there is a vowel on one side of the card then there is an even number on the other side of the card." The correct response is to test that the reverse side of the A is a 4, and that the reverse of the 7 is not an A.

2 Green also noted that these findings have important implications for teams, such as those in which AAW officers work. In groups, he says, "data and warrants may be widely distributed but effective decision-making requires that they be brought together, and arguments for alternative decisions and courses of action constructed and resolved."
indirectly improved communications and coordination behaviors, such as information filtering, proactive communications, and net discipline.

The effects of training in argument are apparently enduring over the long term. Longitudinal research by Kuhn (1991) determined that college-educated individuals were more likely than those with less education to offer proper arguments (those that distinguished evidence from theory, considered the relevance of evidence, and recognized potential rebuttals and alternative theories). Furthermore, the effect was strongest among students educated in a discipline that focuses explicitly on argument: philosophy.

The field research and laboratory studies cited above indicate that argumentation is a common skill and a core skill among decision-makers, including military tacticians. Decision-makers formulate arguments when making judgments. The accuracy of their decisions appears to be a function of their skills at argumentation. The better they identify conflicting interpretations of events, the more boldly they unveil assumptions, the more brightly they illuminate gaps in their knowledge. The better they attempt to resolve these sources of uncertainty, the more accurate is their situation understanding. Most important, however, is that argument is a natural and familiar means of evaluating the evidence at hand and assessing its meaning. For these reasons, we are interested in argument as a window onto situation understanding.

Here we test applications of LSI that exploit argument, specifically the arguments of Navy officers concerning events in a high-fidelity AAW scenario. We apply LSI for three purposes: (1) to determine whether a metric based on arguments can be used to assess overall expertise of individuals, (2) to understand the concepts that constitute situation understanding in specific scenarios, and (3) to learn whether it is possible to automatically diagnose deficiencies in the situation understanding of individuals.

**METHODOLOGY**

**Subjects**

The data analyzed here were drawn from a series of four experiments that tested the effectiveness of training in critical thinking skills. Some 201 active duty military officers participated as subjects while in training programs or other commands of the U.S. Navy. The arguments generated by 31 officers were selected for analysis here because the officers had either unusually great domain experience, or unusually little. Officers in the low expertise group had less than twelve weeks of CIC experience and had not received training in critical thinking skills (which was the focus of the experiments in which these data were gathered; see below) when they provided responses to test questions on a test scenario. These officers had an average of 2 weeks of CIC experience. Officers in the high expertise group had more than twelve weeks experience and they had received training when they executed the selected scenario. These officers had an average of 235 weeks of CIC experience. In sum, the subject groups were quite distinct with respect to the amount of experience and the level of explicit training in critical thinking skills.
Experimental procedure

Experimental designs and procedures varied somewhat between the four individual experiments. Generally, though, participants received training on a high-fidelity AAW simulator, then performed a pretest in which they assessed the intent of experimenter-specified aircraft (tracks) during a tactical scenario. They then received either a control treatment or critical thinking training, in which they practiced generating, critiquing, and refining assessments of track intent by applying instruction in story-building and in identifying and lowering uncertainty. Controls executed the same practice scenarios as officers in the training condition, but did not have the putative benefit of structured training in critical thinking.

During each test scenario, officers answered questions designed to elicit elements of argument: conclusions (i.e., assessments of track intent), evidence for those conclusions, and rebuttals. Specifically, officers were given the following questions (among others), to which they responded in writing in 4-5 minutes each, on average.

- Write down the one assessment of the intent of \{a track indicated by the experimenter\} in which you are most confident. How would you argue the case for this intent?
- State one assessment of intent that is inconsistent with the one in which you are most confident. Assume for the moment that you believe the "inconsistent" assessment is correct. How would you argue the case for this intent?
- Assume that you believe that this track is attacking. (a) Does any evidence seem to conflict with this intent? (b) How would you argue the case for this intent?

Text selection and preprocessing

Responses concerning the final event in one test scenario were selected for the present analysis. That event concerned one of a pair of fighter aircraft that had emerged from enemy territory, apparently attempted to hide beside a commercial airline on a tangent to ownship, and then turned inbound on a possible attack run. The recent presence of a surveillance aircraft suggested that the incoming fighters might have adequate targeting information to attack. The tracks’ vector would take them over two U.S. vessels, ownship and a U.S. warship that had been crippled by a mine only a short time before. However, the aircraft were flying too high to launch missiles at ownship, were arguably flying too slowly for an attack, had not turned on their own fire control radar, and had made themselves conspicuous targets for ownship. Thus, the evidence for a coordinated attack was mixed.

A typical response to one of the questions above, was this 8-argument defense of the assessment that the track was on an attack mission:

- No valid mode 4
- Hostile point of origin
- Political situation indicates possible attack
• It is the weapon of choice for attack
• Targeting assets available
• Hostile profile
• Operation in pair
• Course change for surprise

Responses were transcribed and parsed into arguments at bullets or at punctuation marks. The 31 participants generated a total of 377 arguments.

As expected and confirmed in initial tests, the brevity of the arguments (possibly in combination with the small size of the corpus) hindered LSI in identifying synonyms and important relationships between terms and arguments. To compensate for this, several steps were taken.

A simple stemming routine\(^3\) was run to standardize terms (e.g., "flying" became "fly", "approaches" became "approach"). This improved measures of term similarity. In addition, a small thesaurus was constructed\(^4\) and applied to standardize the variety of spellings of domain terms (e.g., "stbd" became "starboard", "kt" became "knots"). Arguments that cited specific domain terms or scalar values were expanded using the thesaurus to reference more generic entities. For example, "F-4" was expanded to read "F-4 tactical aircraft", and "17,000 ft" was expanded to read "17,000 feet above15k".

As is common in applications of LSI, non-content bearing words (the, him, of) were removed from the texts during preprocessing. However, many of the arguments in this particular corpus focused on spatial relationships between ownship and a suspect track. To retain these relational concepts, the standard list of excluded terms was modified so that prepositions (e.g., between, above, beside) were not stripped away.

Following this pre-processing, an SVD of the corpus of 377 arguments was computed using Bellcore's LSI software. A 75-factor solution was specified because it produced the most accurate correlations with overall expertise (as described below).

**ASSESSING GLOBAL DOMAIN EXPERTISE**

In the first analysis of these argument data, we tested whether the overall expertise of individuals can be assessed using a model based on characteristics of brief arguments. This work extends Landauer's successful experiments with much longer documents (e.g., of 200 words, Rehder, et al., 1998).

---

\(^3\) The routine was an implementation of M.F. Porter's "Algorithm for Suffix Stripping," (Program 14 (3), July 1980, pp. 130-137.) It was implemented in Perl by Ian Phillipps (Copyright Public IP Exchange Ltd (PIPEX)), based on a C version by B. Frakes and C. Cox (1986).

\(^4\) In developing the preprocessing tools described below, we made use of Inferule, a tool developed by Bryan Thompson of CTI that creates a rule base for binary classification of texts, given the texts and the classes (e.g., expert, novice; range, bearing, radar) to which they are known to belong. By studying texts that cannot be properly classified, we were able to specify terms that should be standardized.
Representation of cases of arguments by individuals

The native, LSI representation of each argument in the corpus was a vector of factor weights. For each officer, we constructed two summary vectors, respectively representing the central tendency and variance of all of the officer's arguments. The first summary vector consisted of the mean of factor weights per factor over all argument vectors, per officer. One complication with using the average of factor weights alone is that it does not account for variance between vectors. Consider two extreme cases. One is the case in which all weights on a given factor equal zero. The other is the case in which the sum of the absolute value of negative weights equals the sum of positive weights, i.e., negative and positive weights perfectly counterbalance each other. These cases differ with respect to their variance, but not their average. To discriminate between these cases, we computed a second vector consisting of the standard deviation of factor weights per factor over all argument vectors per officer. Thus, the situation knowledge of each officer was represented by two vectors: one was the central tendency of factor weights over all arguments, the other the variation in factor weights over all arguments.

It is important to note here that the factor vectors used in these calculations were unscaled document factor vectors. It is typical in LSI studies to multiply document factor vectors by a scaling vector that represents the amount of variance in the corpus that is accounted for by each factor. Unscaled factor vectors represent documents (or terms) on the surface of a hypersphere. Scaling effectively projects them off the surface of the hypersphere. This scaling is useful when the bases of indexing and analysis are the same, as when essays are indexed and individual expertise is assessed on the basis of essays. However, unscaled factor weights can be better predictors of variance between groups of authors when the basis for indexing differs from the basis for analysis. (Mike Berry, personal communication). In the present instance, the basis for indexing was brief arguments, but the basis for analysis was the individual's corpus of arguments, or case. The scaling vector (eigenvalues) in this situation represented variance attributable to term distribution over arguments, and not over cases. Furthermore, using normalized, or unscaled document factor vectors potentially allows new scaling functions to be computed for the analytic basis without concern for the biasing effects of scaling on the indexing basis.  

---

5 This problem of indexing on one basis and analyzing on another may be compounded when the indexed corpus consists of extremely heterogenous items (arguments), but when the basis for analysis (cases) is highly homogenous. For example, the arguments in this corpus individually concerned altitude, range, bearing, and so forth. Most officers made arguments on most of these topics. Thus, while their arguments were heterogenous, their cases (of arguments) were homogenous. Altitude was a particularly common topic of argument. Factors that concerned altitude were quite large when scaled by the eigenvalues. Several of the ten factors that accounted for the greatest variance in the text contrasted altitude with other terms. Scaling the argument factor vectors emphasized these altitude arguments. This dimensioned the argument space well. However, high- and low-expertise officers did not reliably differ on altitude arguments. High-expertise officers did differ reliably from low-expertise officers in their references to radar (localization) issues (as discussed below). This topic was prominent only in factors that had lower scaling values. Using unscaled factor vectors allowed these subtle factors to have their due influence as predictors of expertise using the technique described here.
A critical issue in successful use of LSI indexes is selecting the correct number of factors. The corpus used here consisted of 265 terms in 377 documents. There was the potential for LSI to generate a solution with as many as 265 factors. We explored solutions of 50, 75 and 100 factors\(^6\). The proximity computations, described below, were conducted separately for each of these solutions.

**Proximity**

The cosine was computed to measure similarity\(^7\) between each pair of non-identical vectors of mean values (called *cosines of means (MN) vectors*, below), and each pair of non-identical vectors of standard deviations (called *cosines of standard deviation (S.D.) vectors*, below), over all participants. For each participant, we then computed the average of cosines of means vectors over all members of the low expertise group. This produced a measure of the similarity of the individual to the "average" low expertise officer. The average of cosines of MN vectors to the high expertise group was computed, and similar averages were computed using the S.D. vectors for each group. Thus, each individual was represented by four values: a mean of cosines of means vectors with respect to the low expertise group, a mean of cosines of S.D. vectors with respect to the low expertise group, and a similar pair of cosines with respect to the high expertise group.

A close inspection of the distribution of values on these variables indicated that, for the 75- and 100-factor solutions, kurtosis and skewness were high in the case of the average of cosines of S.D. vectors for the low group. These variables were transformed using the recommended reflection and inversion (Tabachnick and Fidell, 1996). (We also report, below, the results for the 75- and 100-factor solutions using untransformed variables.)

**Predicting Overall Expertise**

A discriminant analysis was performed for each of the 50-, 75-, and 100-factor solutions. The grouping variable (the variable to be predicted by the discriminant functions) was expertise level (high vs. low). Predictors were the four mean proximity values defined above. All four averages were statistically reliable predictors of level of expertise in all of the solutions reported here.

---

\(^6\) Solutions of low rank are abbreviations of higher rank solutions. Thus, the first 50 factors were identical in all solutions, and the first 75 factors were shared by both of the larger solutions.

\(^7\) Several metrics are commonly used to summarize the relative and absolute locations of points in multidimensional space. The cosine is by far the most common measure of proximity in LSI experiments, and has produced statistically reliable results in studies correlating LSI and SME judgments of expertise from text (Rehder, et al., 1998). In a mathematical analysis of several proximity measures Rehder presents evidence that the cosine can be interpreted as a measure of situation-specific knowledge. In contrast, the length of a vector appears to represent somewhat more general knowledge of the domain. The dot product, not tested here, can be interpreted as the interaction of the cosine and length. However, it is not necessarily a useful metric of similarity. Rehder found that the dot product did not reliably predict expertise scores when added to regression models already containing cosine and length.
The quality of the solution was judged by the precision and recall of Jackknifed 
classification. Precision and recall are widely used performance measures in 
information retrieval. In the present context, precision denotes the ratio of the number of 
officers who are correctly predicted (by the discriminant function) to belong in a category 
to the total number of officers predicted to belong (correctly or incorrectly) in that 
category. For the designer of an instructional system, low precision rates are a concern 
because they indicate that many individuals whom the model predicts belong in one 
category in fact belong in another. Recall is defined as the ratio of the number of officers 
correctly predicted to belong in a category to the total number of officers who in fact 
belong in that category. Low recall rates suggest that the model fails to place many 
individuals who belong in a particular category into that category. In the Jackknifed 
approach, a unique set of discriminant functions is generated to predict the class 
membership of each individual, and those functions are generated using all of the data 
except the individual's own data. Thus, the prediction is not biased by the individual's 
own performance. This replicates field application in that the individual is classified on 
the basis of the performance of peers who have previously been evaluated.

As illustrated in Figure 1 (see data point A), overall accuracy was best using a 75 
factor solution with a transformed average of cosines of S.D. vectors for the low 
expertise group. In that instance, mean precision was 89.5% at a mean recall rate of 
88.5% ($F_{4,26} = 5.719, \ p = 0.0019$). Figures for the high-expertise officers were 100% 
precision and 75% recall; for the low-expertise officers, precision was 79% and recall 
was 100%.
Figure 1. Precision and recall results for LSI-based expertise assessment.

Note:

<table>
<thead>
<tr>
<th>Plot symbol</th>
<th>Factors</th>
<th>Predictors</th>
<th>Mean precision</th>
<th>Mean recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75</td>
<td>Mean &amp; SD*</td>
<td>89.5%</td>
<td>87.5%</td>
</tr>
<tr>
<td>B</td>
<td>75</td>
<td>Mean, SD*, &amp; length (overlaps E)</td>
<td>81.1%</td>
<td>80.8%</td>
</tr>
<tr>
<td>C</td>
<td>75</td>
<td>Mean, SD*, &amp; arguments</td>
<td>77.5%</td>
<td>77.5%</td>
</tr>
<tr>
<td>D</td>
<td>75</td>
<td>Mean &amp; SD</td>
<td>82.7%</td>
<td>81.0%</td>
</tr>
<tr>
<td>E</td>
<td>100</td>
<td>Mean &amp; SD (overlaps with B)</td>
<td>81.1%</td>
<td>80.8%</td>
</tr>
<tr>
<td>F</td>
<td>50</td>
<td>Mean &amp; SD</td>
<td>71.8%</td>
<td>71.2%</td>
</tr>
<tr>
<td>G</td>
<td>75</td>
<td>Mean &amp; SD**</td>
<td>63.3%</td>
<td>62.8%</td>
</tr>
<tr>
<td>H</td>
<td>75</td>
<td>Mean &amp; SD</td>
<td>60.2%</td>
<td>60.0%</td>
</tr>
</tbody>
</table>

SD* denotes the use of transformed average of cosines of S.D. vectors for low expertise participants. SD** indicates the use of transformed average of cosines of S.D. vectors.
for low- and high-expertise groups. Group definitions for G and H exclude officers with training in critical thinking (see full definition, below).

These results were not improved upon by adding two promising predictors: vector length and argument count. Vector length, as indicated above, can be interpreted as an index of domain-specific, but not context-specific knowledge. In this sample, the length of the vector of mean factor weights was a nearly significant predictor of expertise level in a t-test ($t_{27.7} = 2.061$, $p = 0.054$), but its use decreased the accuracy of jackknifed classification tests slightly to a mean of 81% for both precision and recall. Argument count typically rises with critical thinking training, and it was a reliable predictor of expertise level among these participants ($t_{27.7} = 2.603$, $p = 0.015$). However, overall accuracy declined to 78% for both precision and recall with the use of argument counts. Both vector length and argument count were discarded as predictors.

Using untransformed averages of cosines of S.D. vectors for the low expertise group, the 75-factor solution produced mean precision of 83% and mean recall of 81% ($F_{4,26} = 5.216, p = 0.0032$). Precision and recall figures for each expertise category were five to eight points lower than in the best solution.

The accuracy rate for transformed and untransformed data using the 100-factor solution was 81% for precision and for recall ($F_{4,26} = 5.384 p = 0.0027$), a figure that was well above the average rates for the 50 factor solution: 72% mean precision and 71% mean recall ($F_{4,26} = 5.371, p = 0.0027$).

We compared these results to those from a group of participants whose expertise was defined in a manner that eliminated the effects of training in critical thinking skills. The only responses analyzed were from participants on pretests or in the control condition. The definition of this group also emphasized CIC experience.

- All low-expertise officers had some CIC experience, but no experience in AAW or TAO positions, and no training in critical thinking for tactical decision making. Officers in this group had a mean of 37 weeks experience in the CIC, as opposed to a mean of 2 weeks for the group described above.
- High-expertise officers had experience in AAW or TAO positions, and no training in critical thinking for tactical decision making. Participants in this group had 97 weeks experience in the CIC, on average, vs. 235 on average, above.

Using a 75-factor solution, category membership for officers was predicted with 63% mean precision and 63% mean recall, provided that averages of cosines of S.D. vectors for low and high expertise groups were reflected and inverted to correct distribution problems. Mean precision and recall rates were 60% without transformation. Both figures are close to chance (50%) on this dichotomous classification task; accordingly, neither discriminant function produced statistically reliable results.

In sum, these analyses demonstrate that, given a strongly bimodal sample, the expertise of CIC officers can be predicted with 87% accuracy from their free text arguments concerning assessments of track intent. For groups that are more homogeneous in terms of CIC expertise and prior training in tactical decision making, predictions of expertise level approach chance.
MODELING DOMAIN KNOWLEDGE

As a step towards understanding specifically how the knowledge of more- and less-expert officers differed, we attempted to interpret the factors generated by the index of their arguments. To facilitate interpretation, the indexed arguments were manually coded into 15 categories developed in prior analyses of intent assessment by highly experienced Navy officers (Cohen and Freeman, 1997; Freeman and Cohen, 1996). Subcategories of approach concerned aspects of track kinematics. In particular, arguments classified as approach: commercial air route concerned the suspicious use of a commercial air route by the track being assessed, while approach: course addressed more general comments concerning the route. Arguments categorized as approach: profile were the most general of all approach arguments; they were global statements about track kinematics such as "outside commercial air profile" or "not committed to hostile flight profile." The background category was typified by references to recent events that might weigh on the plausibility of a particular assessment. (For example, a recent air strike by America against a hostile nation would make it more plausible to conclude that a suspect aircraft was on an attack mission against American vessels.) Arguments concerning capability addressed the ability of the platform to carry out a specific mission (such as an attack or search and rescue). Origin and destination arguments speculated on the endpoints of the track’s route. Response arguments concerned response (or lack of response) to IFF and to verbal warnings to change course away from ownship. The few arguments concerning opportunity addressed the presence of an opportunity for attack (e.g., ownship) or for surveillance (e.g., a nearby mine-damage repair operation). Localization arguments concerned evidence or inferences about the track’s use of radar, its ability to do so, or the implications of radar type for identifying the track. Arguments that were unique (i.e., unclassifiable) or uninterpretable were labeled miscellaneous.

Multi-dimensional scaling (MDS) was used to reduce the dimensionality of the factor space and improve both the proportion of dimensions that could be interpreted and the meaningfulness of those dimensions. The cosine between every pair of non-identical factor vectors was computed to create a table of proximities between arguments⁸. An MDS solution computed using these proximity data produced five dimensions⁹. An examination of the distribution of arguments by dimension suggested that there were dense clusters of arguments in one or a few categories at each pole in the midst of clouds of arguments that otherwise consisted of a relatively equal mix of arguments in other categories. To clarify interpretation, we attempted to strip away these clouds. On each dimension, four hinges were computed between the argument with the largest positive coordinate and the argument with the largest negative address. Each outer hinge demarcated the 20% of arguments with the most extreme weights (those that lay closest to a pole). The distribution of arguments by category within those extreme regions was then computed. This distribution was normalized with respect to

---

⁸ Unscaled factor vectors were used, as in the previous analysis.

⁹ Kruskal and Guttman loss functions produced solutions with the same dimensionality. The Kruskal solution is interpreted here.
the relative frequency of arguments in each category\textsuperscript{10}. Thus, relatively rare arguments, such as opportunity and capability were heavily weighted, while common arguments, concerning course or altitude, for example, received lighter weights.

The five MDS dimensions were characterized by surprisingly strong weights on altitude (dimension 1); approach profile and response to warnings and IFF interrogation (dimension 2); altitude, origin, destination, speed, and use of a commercial air route (dimension 3); capability, range, destination, formation, and speed (dimension 4); and localization (dimension 5) (see Table 1).

This distribution of classes of arguments suggested four psychological constructs, or \textit{elements} of situation understanding: track response, track route, track kinematics, and track localization capability. The track response element was represented by arguments concerning responses to verbal warnings and IFF interrogation. The route element was defined by arguments concerning track altitude, origin, destination, and adherence to a commercial air route. The kinematics element was represented by range, speed, and formation (see comments on interpretation, below). The localization element was characterized by arguments concerning the track’s use of localization systems or capability for localization. All remaining argument categories (course, profile, background, opportunity, capability) were considered to be part of a miscellaneous element.

The definition of these elements did require interpretive judgement. In particular, destination arguments were employed only in the route element, where they were prominent, and not in the kinematic element, where they were also concentrated. Speed arguments were placed in the kinematic element, though they were not prominent on the dimension that otherwise defined that element. Capability was dropped from the kinematic element because it was not meaningful there, and because it’s inclusion in this element reduced overall classification accuracy (below). Even considering these adjustments of three of the 15 argument topics, interpretation was reasonably straightforward and rapid.

Table 1. Dimension Definitions Using Weighted Categories

\footnotesize 10 The normalization formula was the reciprocal of the number of arguments in that category over all arguments in all categories.
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Argument categories with high weighted frequency in the dimension</th>
<th>Model element</th>
<th>Argument categories constituting model element</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>altitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>approach profile, response via communications or IFF</td>
<td>Response</td>
<td>response via communications or IFF</td>
</tr>
<tr>
<td>3</td>
<td>altitude, origin, destination, speed, commercial air route</td>
<td>Route</td>
<td>altitude, origin, destination, commercial air route</td>
</tr>
<tr>
<td>4</td>
<td>capability, range, destination, formation, speed</td>
<td>Kinematics</td>
<td>range, formation, speed</td>
</tr>
<tr>
<td>5</td>
<td>localization</td>
<td>Localization</td>
<td>localization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Miscellaneous</td>
<td>All remaining categories (course, profile, background, opportunity, capability)</td>
</tr>
</tbody>
</table>

In sum, situation understanding for this particular AAW problem was represented by five elements: track response, track route, kinematics, localization, and a miscellaneous element.

**DIAGNOSING DEFICIENCIES IN DOMAIN KNOWLEDGE**

The elements or constructs, above, were illuminating, but not necessarily highly useful. Two analyses of their utility were conducted. The first evaluated whether arguments could be accurately classified into elements automatically. If this could be done, then an instructional system could be designed to automatically grade the completeness of cases of arguments or rate the distribution of arguments across elements. The second analysis attempted to pinpoint differences in the distribution of arguments across elements between low- and high-expertise groups.

Support for Automatic Classification of Arguments

A test of the feasibility of automatically classifying arguments into elements was conducted in the following manner. The proximity (cosine) of every pair of non-identical arguments was computed using unscaled factor vectors. For each argument, the mean proximity was computed between arguments within each of the five elements. For example, for the route argument "He is flying at 18000 feet", we computed the average proximity to all other arguments concerning route, as well as those concerning response, kinematics, localization and the miscellaneous element.

Discriminant functions were computed in jackknife fashion using the five means (one per element) as predictors of argument membership in elements. Over all model
elements, average precision\textsuperscript{11} was 84\% and average recall was 77\% ($F_{20,1208} = 105.776$, $p < 0.0001$). Precision rates ranged between 84\% and 100\% for all elements except miscellaneous (59\%). Recall rates ranged between 67\% and 88\% for all elements\textsuperscript{12, 13, 14}.

\textsuperscript{11} In this context, precision denotes the ratio of arguments that are correctly predicted to belong in an element to all arguments correctly and incorrectly predicted to belong in that element. To the designer of an instructional system, low precision indicates that the automated classifier places many arguments in an element that belong elsewhere. Recall is defined as the ratio of arguments that are correctly predicted to belong in an element to all arguments that belong in that element. Low recall suggests that many arguments that belong in an element are mis-classified.

\textsuperscript{12} The low precision rate in the miscellaneous category was expected. Discriminant models assign difficult-to-classify cases to the category with the greatest variance, or dispersion. That the recall rate was slightly higher than the precision rate suggests that the arguments constituting the miscellaneous element lie fairly close to one another in the multidimensional space defined by LSI. Nonetheless, an attempt was made to divide the miscellaneous category into more distinctive categories, which might raise precision and recall. Three types of arguments — background, opportunity, and capability — are high-level issues more often considered by supervisors and commanding officers than by team members with lesser responsibilities. The remaining two argument categories, course and profile, seem to constitute a (second) route element. Thus, miscellaneous arguments were split into a command element and second route element. In this classification scheme, mean recall and mean precision were 46\% ($F_{30,1450} = 13.208$, $p < 0.0001$). Furthermore, precision was quite low for the command element (25\% at 51\% recall) and the second route element (40\% at 50\% recall), which was well below the precision rate of 76\% at 49\% recall for the original miscellaneous element. This restructuring of the model elements was rejected. An alternative approach was to reconceptualize this group of arguments. No label unifying these arguments could be conceived, however. Pending the discovery of such a concept, this element will continue to be called "miscellaneous."

\textsuperscript{13} In a separate analysis, we tested the accuracy of classification based solely on the maximum average proximity value. For example, an argument with proximity ratings of .35 for kinematics and .55 for localization was classified as a member of the localization element in this analysis. Average precision was unacceptably low at 71\%, though recall rose slightly to 80\%. Precision for the localization element suffered most, dropping 20 points to 66\% at 94\% recall.

\textsuperscript{14} An analysis in which capability was included in the current kinematics element produced overall recall rates of 45\%. This suggests that arguments concerning capability are spatially distant from those concerning range, destination, formation, and speed.
Figure 2. Precision and recall results for LSI-based model element categorization.

Note:

<table>
<thead>
<tr>
<th>Plot symbol</th>
<th>Model element</th>
<th>Mean precision</th>
<th>Mean recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Route</td>
<td>84%</td>
<td>67%</td>
</tr>
<tr>
<td>B</td>
<td>Kinematic</td>
<td>93%</td>
<td>81%</td>
</tr>
<tr>
<td>C</td>
<td>Localization</td>
<td>86%</td>
<td>88%</td>
</tr>
<tr>
<td>D</td>
<td>Miscellaneous</td>
<td>59%</td>
<td>80%</td>
</tr>
<tr>
<td>E</td>
<td>Response</td>
<td>100%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Differences in Situation Understanding by Level of Expertise

Templates for high- and low-expertise groups differed reliably in the frequency with which they made arguments concerning exactly one model element: localization ($\chi^2_1 = 3.840, p = 0.05$). Officers in a high-expertise condition generated approximately twice as many arguments concerning localization as did officers in the low-expertise group.
The elevated attention to localization issues here was consistent with prior findings from empirical and field studies. Cohen and colleagues (Cohen and Freeman, 1997; Freeman and Cohen, 1996) found that training in critical thinking skills increased the frequency of arguments concerning localization. In qualitative analyses of the TADMUS interviews, Cohen and colleagues (Cohen, Freeman, et al., 1994) found localization to be a surprisingly frequent concern among experienced officers, one that heavily influenced the reading of more standard kinematic cues. For example, officers tend to regard as hostile a contact that emerges from hostile waters, turns toward own ship, and accelerates. But if this contact is too far away to have detected own ship, these cues must mean something else. If hostile intent is still suspected, then the officer may look for third-party targeting platforms or discuss with intelligence assets the possibility that the track carries unusual localization equipment. Conversely, in other TADMUS incidents, when a track was too slow or too high to fit a hostile profile, its behavior was sometimes reconciled with a hostile intent story by assuming that it was attempting to localize a target.

The current findings differ from prior results with respect to attention to kinematic issues, however. Cohen, Freeman, et al. (1994) found that officers in a high expertise condition (those who had received critical thinking training) generated a lower proportion of kinematic arguments than did officers in a lower expertise condition. Officers in the high expertise group analyzed here generated almost precisely the same proportion of arguments in the kinematic element (82%) as did officers in the low expertise group. One factor that might explain the discrepancy with prior findings is that kinematic arguments were not confined to a single element (the kinematics element). We tested whether groups differed on kinematics arguments, defined as the union of the kinematics element and the route element. The groups remained statistically indistinguishable. We conclude either that reference to kinematic issues is not a reliable correlate of expertise, or that the current model does not support the same definition of kinematic issues; it is not possible to partition the current model at element borders into classes of kinematic and non-kinematic arguments similar to those used in prior research. For example, course, a kinematic issue, occupied the miscellaneous element with a number of non-kinematic topics.

**DISCUSSION**

We have taken argument as a window onto situation understanding. We have applied Latent Semantic Indexing to a corpus of arguments to assess domain expertise globally, to help define cognitive models, to diagnose situation understanding in detail. Applying LSI to a officers’ arguments produced highly accurate assessments of their domain expertise. Overall level of expertise in this complex domain was predicted from a summary of arguments with 89.5% precision at 88.5% recall. Reducing the LSI factor space for this argument corpus proved useful in defining five elements of situation understanding in this scenario. Officers at two levels of expertise reliably differed in their attention to one of these elements: localization. These differences in situation understanding can be automatically diagnosed using a method explored here. It classified specific arguments into model elements with 84% precision at 77% recall overall. For the element that distinguished levels of expertise (localization),
Categorization reached 86% precision at 88% recall. These findings have implications for cognitive modeling during task analysis, for developing student models, and for assessing student performance more generally.

Cognitive Modeling

In well-defined domains, such as mathematics, cognitive task analysis can largely be conducted by examining textbook solution strategies, and through careful observation of human problem-solving (Ginsburg, 1989) and analysis of problem solutions (Jones and van Lehn, 1994). In ill-defined domains, however, problem solving often involves a complex and dynamic orchestration of recognition (or pattern-matching) and metacognition (Cohen and Freeman, 1997) that is commonly called judgement. Often, neither the processes nor products of problem solving are reliable between practitioners (Shanteau, 1999; Shanteau and Stewart, 1992), nor are they necessarily easy to observe. To perform cognitive task analysis in such domains, researchers have often applied verbal protocol analysis. Critical incident interviews (Flanagan, 1954; Weitzenfeld, et al., 1991), for example, are often used to elicit highly detailed descriptions of problems that challenge expertise, and the domain knowledge and problem-solving processes brought to bear on them. Collecting and analyzing interview data is laborious, time-consuming, and expensive. It is often difficult and expensive to achieve high inter-rater reliability. Furthermore, in qualitative analyses that involve categorization, there is also the murky problem of how analysts develop a coding scheme. However, the manual approach is often deeply informative.

We do not suggest that the modeling technique tested here is a replacement for verbal protocol analysis. However, the automated method has the benefits of being rapid, highly systematic, and transparent. Automated text preprocessing, LSI, and the various statistical techniques applied produce replicable results, the techniques are documented in detail, and the data can often be made publicly available. There is some judgement in the automated method, just as in manual protocol analysis. Interpreting the factor space as a set of elements involves evaluating alternative interpretations of dimensions. However, the room for variance between analysts is not large, and the data on which analysts may disagree is easily summarized (see Table 1) for peer review.

It might be argued that results similar to those obtained here can be got with card sorts, comparative rating tasks, categorization tasks, and related laboratory techniques for identifying knowledge structure. We find this to be a dubious argument for several reasons. First, these techniques are patently artificial compared to the analysis of argument, particularly when argument arises naturally in the course of task execution. Second, a small set of terms is typically used in these tasks, which seems to assume that the content of domain knowledge is small or focused on a few critical concepts. Both of these are dubious assumptions. Argument-based tasks do not constrain language use, thus they do not artificially constrain the concepts that are available for analysis. Third, these laboratory tasks necessarily employ a standardized set of terms as stimuli. This assumes that the contents of domain knowledge are constant between domain practitioners and that only knowledge structure varies. But the contents of domain knowledge does vary between practitioners, and this is a central concern in the
study of expertise. A final objection to these laboratory techniques, unrelated to their validity, is that they require that a task analysis be performed in order to develop the list of terms to be categorized or sorted. The LSI approach does not.

The laboratory tasks mentioned above and the LSI-based method explored here do share one shortcoming: neither reveals much about knowledge processing. However, there is the potential to modify LSI algorithms to take into account the order of arguments (Psotka, personal communication), and this might shed light on the processing that underlies situation understanding.

In sum, LSI is a promising tool for improving qualitative analysis of human knowledge of complex tasks. It is no substitute for manual verbal protocol analysis, but it has its place. At the least, the technique may guide researchers in their manual analyses. At best, the LSI-based approach may help researchers to validate such analyses, or extrapolate manually developed models in new subdomains.

Student Modeling and Assessment

Independent of its value or validity as a tool for qualitative analysis of situation understanding, LSI has great potential for assessing student knowledge. In this study, we paired LSI with discriminant analysis to classify arguments into elements of a cognitive model. It was irrelevant that the model was developed using LSI. Sorting tasks, verbal protocol analysis, and other techniques are perfectly suitable sources of cognitive models. The diagnostic technique presented here can be viewed as a test of such models. In these tests, we use LSI and discriminant analysis to discern whether (1) each element of the cognitive model is instantiated by some argument and (2) the classification of arguments into elements is relatively accurate. Uninstantiated elements are useless appendages to a model; they indicate invalid constructs. Low classification accuracy with respect to an element indicates that the element is associated with a range of arguments that have little to do with each other. When a large test set of arguments can be classified with relatively high accuracy, and when all elements of the model are instantiated there is reason to trust that the model is psychologically valid, and that LSI can leverage the model to classify arguments and diagnose performance.

Argument classification has a number of applications. The distribution of arguments by a student over model elements is arguably an indication of the student’s understanding of the given problem. For example, when a student fails to cite localization in an AAW scenario or mentions it rarely (relative to domain experts), there is reason to believe that the student does not deeply understand the role of localization in AAW operations. Reliable differences in the distribution of arguments over elements between a student and experts should trigger remediation in a highly interactive learning environment. Such remediation might consist of a hint, such as a reminder about specific evidence concerning localization, explicit instruction concerning the importance of localization in AAW, presentation of training scenarios that hinge on localization issues, and/or tests to validate deficiencies regarding this topic.

A more subtle application of the classification technique is to use the pattern of element instantiation (i.e., the order in which arguments are made, as well as argument
content) as an indicator of cognitive processing. This is a form of student modeling that might predict subtle gradations of expertise, or that might explain specific decision errors\textsuperscript{15}. This is a substantial research topic that we are only beginning to address.

Recall that two levels of assessment were explored here: global prediction of expertise and a fine-grained evaluation of situation understanding. We are concerned that the approach to assessing global expertise was useful for discriminating between groups of practitioners who differed strongly in their domain experience, but not between less distinctive groups. We hope to explore this limitation and ways of ameliorating it. For example, this may be a less pronounced issue in richer domains. AAW operations involve a surprisingly small variety of cues, assessments, and responses in contrast to political science, history, and literature, for example. It may also be possible to improve global assessment by processing a broader range of arguments. Recall that this study concerned only arguments supporting assessments and arguments concerning conflicting evidence. Arguments concerning missing information and weak assumptions are often voiced and might be correlated with subtle differences in expertise. Finally, there may be better definitions of the levels of expertise in AAW than the ones used here, and that, too, would presumably influence the accuracy with which these methods distinguish differences in domain expertise. Note, however, that the issue of a strongly bimodal subject sample is irrelevant to argument classification. The relatively high accuracy of argument classification in this study was obtained over all members of the sample without reference to expertise.

We conclude with an architectural metaphor for the application of LSI to arguments. Consider a home in the colonial style. Each window consists of several panes separated by a gridwork of white muntins. From outside, the window as a whole offers a view into the privacy of the home. Likewise, arguments offer a view into the private space of the mind. LSI, like the muntins, organizes that view. To the extent that the view through each pane is distinctive and meaningful, this organization potentially defines useful constructs. The work presented here suggests that LSI can indeed be used to identify such constructs for purposes such as task analysis, however, the techniques used here must be refined and validated in other studies. Independent of this, we have shown that LSI can leverage the latent semantic content of short arguments to diagnose situation understanding in detail, and to assess domain expertise more globally.

###

\textsuperscript{15} For example, Adelman et al. (1997) have demonstrated that errors in Army AAW operations are in part a function of the interaction of the order in which evidence is presented (recency and primacy effects) and the story (or framing) with which Army AAW officers interpret evidence. To the extent that arguments reveal both evidence and framing, a real-time analysis of the order and content of arguments might reveal decision errors in the making, in time for rapid remediation in an instructional system or a decision support device.
REFERENCES


**ACKNOWLEDGEMENTS**

This work was funded by the Navy Personnel Research & Development Center and supported by Bellcore through the loan of LSI software.