Integrating Knowledge-Based and Case-Based Reasoning

Timur Chabuk
Department of Computer Science
University of Maryland
College Park, MD 20740
chabuk@cs.umd.edu

Abstract: There has been substantial recent interest in integrating knowledge based reasoning (KBR) and case-based reasoning (CBR) within a single system due to the potential synergisms that could result. Here we describe our recent work investigating the feasibility of a combined KBR-CBR application-independent system for interpreting multi-episode stories/narratives, illustrating it with an application in the domain of interpreting urban warfare stories. A genetic algorithm is used to derive weights for selection of the most relevant past cases. In this setting, we examine the relative value of using input features of a problem for case selection versus using features inferred via KBR, versus both. We find that using both types of features is best (compared to human selection), but that input features are most helpful and inferred features are of marginal value. This finding supports the idea that KBR and CBR provide complimentary rather than redundant information, and hence that their combination in a single system is likely to be useful.

INTRODUCTION

In many application fields, expert-level problem solving naturally involves reasoning from a combination of both general knowledge and individual past cases. Well-known examples of this occur in legal reasoning, medical diagnosis and management, military tactical planning, software engineering, and related areas. From the viewpoint of those developing AI systems intended as decision aids, the need for reasoning from both general knowledge and individual cases has led to substantial recent efforts to find ways to integrate these two approaches within a single framework (reviewed in Marling et al, 2002), and this continues to be an active research area today.

In this context, we are investigating the feasibility of creating an application-independent approach to interpreting multi-episode "stories" that combines a variety of AI reasoning methods (rule-based reasoning/deduction, cause-effect reasoning/abduction, Bayesian inference, constraint-satisfaction problem solving, etc.) with the retrieval of past

related cases. Our goal is to implement a system that, given an application-specific knowledge base plus a database of past cases represented in terms of the same features, is able to generate inferences about new situations by concurrently using both knowledge-based reasoning and examination of past cases. This is an ambitious goal that involves addressing a number of challenging issues related to understanding narration [Herman, 2003]. Our central hypothesis is that specific human-readable knowledge descriptions, written in a simple but formal knowledge representation format, contain sufficient information about an application area's ontology and terminology to enable a natural language "story interpretation system" to be generated automatically.

In this paper, we focus on one aspect of such a general system, the issue of whether the inferences made by a knowledge-based reasoning process can help guide the identification of the most relevant past cases during problem solving. Effective retrieval of related cases is widely recognized to be very important to successful applications of case-based systems [Pal & Shiu, 2004] and it is one way in which synergistic integration can occur. More specifically, we focus here on the issue of the relative value of input features of a problem versus inferred features in guiding the retrieval of the most relevant past cases. By *input feature*, we mean an evident/observable aspect of a specific problem that serves as input to a decision aid (e.g., for a medical diagnosis system, a patient's age or a symptom), while *inferred feature* refers to an inference made by the system (e.g., diagnosis, recommended treatment, or prognosis). While our approach is intended to be general in nature, for concreteness and because our most recent attention has focused on this topic, in the following we present our work in the context of a specific application, the understanding of multi-episode stories involving urban warfare. There are many sources of such stories available in natural language format (e.g., [Antal & Gericke, 2003; Grau, 1996; Keegan, 1994]), and it is unlikely that any person can memorize the large volume of information and lessons they contain. Thus, if relevant episodes could be quickly identified by an automated system, they would provide a rich source of potentially useful information for military commanders who must make real-time tactical decisions.

METHODS

Our work on effective case retrieval is being done within the context of a broader study, as follows. An application-specific story interpretation system built within our framework works as illustrated in Figure 1. A user supplies a narrated description of an *episode*, a set of events that have occurred in the application domain. A natural language (NL) parser translates the episode into a set of input features, and from these the inference method(s), such as rule-based deduction, derive various conclusions (inferred features)

using domain specific knowledge. These conclusions, plus the most relevant known past related cases (or episodes), are retrieved.

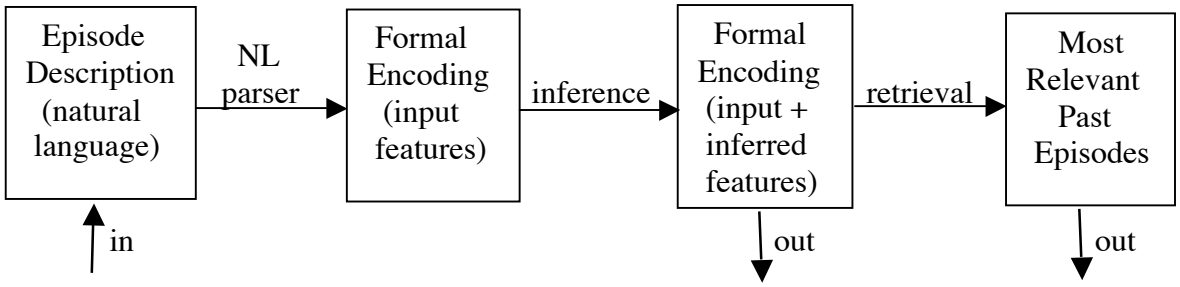

Figure 1: User's view of an application-specific "story understanding system". ■

An application system like that described above (Figure 1) is built by a domain-independent *constructor* as illustrated in Figure 2. The constructor takes two inputs, an application-specific source file ("knowledge base"), written in a simple knowledge representation language, and a database of past multi-episode stories. The constructor uses this information to generate the NL parser, inference mechanisms, and case retrieval software needed in the application-specific story interpretation system (Figure 1), and encodes the cases in terms of the source file's attributes for later retrieval. At present, the constructor is implemented, but more work is needed on the urban warfare source file in order for our keyword parser to be able to parse all of the cases. As a result, the encoding of some cases for the experiments described below has been done manually.

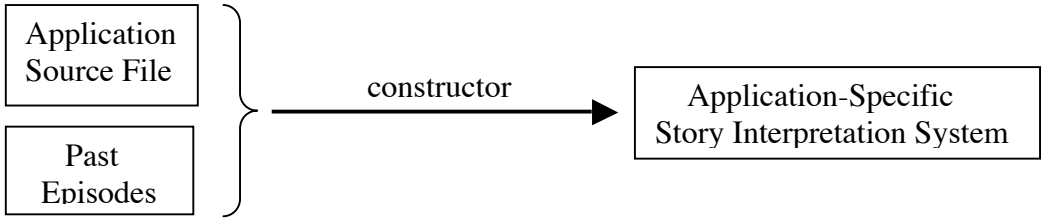

Figure 2: Story interpretation systems like that shown in Figure 1 are constructed automatically from an application-specific source file and a database of past episodes. ■

For example, for the urban warfare domain, the current source file encodes a set of input and inferred attributes, along with their possible values, plus knowledge in the form of production rules and simple descriptions. The attributes form a hierarchy that is implicitly defined by the encoded knowledge. Attributes are defined as being single or multiple-valued. There are forty-three input attributes and thirteen inferred attributes in the existing urban warfare source file. Most knowledge is in the form of production rules, but for two of the inferred attributes simple pattern matching and scoring is used. In some

cases, the inferred features (attribute value assignments) represent abstractions of the input features. Here is an example of an input attribute named "arms" in the source file:

arms [mlt]:
small arms [synonyms: pistols, hand guns, rifles, guns],
light machine guns [synonyms: submachine guns, automatic rifles],
anti-tank weapons [synonyms: rocket propelled grenades, RPG], …

This declaration conveys that "arms" is an attribute/property of a story that, for a specific problem, can simultaneously take on multiple ("mlt") possible values such as "small arms" and "light machine guns". Ontological information is implicitly present in the synonym declarations, so the NL parser (Figure 1) can recognize that the presence of "RPG's", for example, means that (arms = anti-tank weapons) is an appropriate interpretation. An inferred attribute, such as

assessments [mlt]:
enemy likely has outside support,
civilians at great risk, …

is defined in a similar fashion, but has its value determined by the inference mechanism rather than the NL parser, e.g., via rule-based deduction. Rules are defined in terms of attribute values. For example,

IF (enemy organization = militia) AND (enemy technological level = high),
THEN (assessments = enemy likely has outside support)

is a rule in the current knowledge base. This rule indicates that, if an irregular military group ("militia") has high tech weapons, the system should make the inference that the group probably has outside support.

In the current implementation of our system, an abductive keyword natural language parser (Figure 1, on the left) is used to extract the values of the input attributes from stories written in natural language text. As the constructor processes an application-specific source file, each word encountered is indexed to the attribute/value it helps to name. The resultant NL parser processes a subsequent story word by word, and as each word is processed, it evokes the set of all possible senses (assertions about the value of known attributes) of that word. Words can be very ambiguous. For the urban warfare domain, the system evokes five different senses for the word "fire", such as the phenomena of friendly fire or the tactic of setting buildings on fire. The presence of a word in a story is explained in a context-sensitive fashion by making an assertion about an attribute's value. Disambiguation of the word's meaning is done using a parsimonious covering process, a type of abductive reasoning [Josephson, 1994; Reggia, 1992] that

results in a set of assertions representing the interpretation of the story. For example, the presence of the word "fire" may be explained by the assertion (our tactics = set fire to building) or several other assertions.

For the urban warfare scenario used in our experiment below, a case-base was established containing 30 episodes from ten different stories. The stories spanned the time period from just before World War II through the mid 1990's, and were taken primarily from Russia's Chechen Wars [Oliker, 2001], the Wikipedia online encyclopedia [Anon, 2003], and City Fights [Antal, 2003].

An assembled system like that for the urban warfare scenario identifies the past most relevant episodes by measuring the similarity of every episode in the case-base using a linearly weighted distance metric. This *similarity function* takes two episodes and outputs a numeric score indicating their degree of similarity in terms of their attribute values. The contribution of an attribute to this similarity score depends on whether the attribute is single or multiple valued, and whether it is nominal or ordinal. Multiple valued attributes are treated as being a collection of single valued attributes, where each value can be either present or absent. In the similarity function, a positive real-valued weight between 0 and 10 is associated with each single-valued attribute, and with each possible value of multiple-valued attribute. The similarity of two episodes on each attribute are multiplied by their respective weights, and then summed to produce the final overall similarity score. Since the optimal weights for case retrieval are not known a priori by our application-independent system (Figure 2) for a specific application, our approach is to include with each episode in the case database the identity of the other single most similar case that is present. This best match identity is specified by a person at the time of case-base creation and represents the "gold standard" for the evaluation described below.

As others have done [Dubitzky, 2001], we used a genetic algorithm (GA) to automatically evolve the set of weights to be used by the similarity measure in the resultant application-specific system. In the urban warfare example used here, the GA population was 600 haploid chromosomes, each a vector of real-valued weights to be used by the similarity measure. Tournament selection of reproducing parents (with elitism) was used, with probability of double-point crossover 0.35 and of mutation 0.60. If selected for mutation, each real number in the chromosome had a 10% chance of being replaced with a random number between 0.0 and 10.0. The fitness of a chromosome's set of weights was based on how well they correctly identified the a priori human-identified most similar other case in the case base for each and every existing case. The genetic algorithm was run for 300 generations and the most fit set of weights in the final population was used in the application-specific similarity measure.

While the inferred features are intended in and of themselves to be useful to a human operator of the system (leftmost "out" arrow in Figure 1), we consider here whether or not they are also useful when trying to automatically assess similarity of episodes for case-retrieval. To address this issue, we evaluated the ability of the story interpretation system to identify the best match in the case base for input episodes when using all attributes, input attributes only, and inferred attributes only. This allowed us to examine one aspect of the impact of integrating knowledge-based reasoning with case retrieval. A standard leave-one-out strategy was used to evaluate how well our methods for evolving attribute weights with specific data generalize. The episodes from one story were removed from the training data and an optimal set of weights evolved using the rest of the episodes in the case-base as training data. That set of weights was used to find the most similar episodes to the episodes that had been removed, and for each of the removed episodes we recorded how highly the system ranked its actual most similar episode (according to the "gold standard"). Attribute weights for the similarity measure were independently evolved using all of the attributes, only the input attributes, and only the output attributes. These three experiments were repeated 10 times each, each time excluding the episodes from a different story from the training-data. We also performed the three experiments while leaving out no story episodes in order to establish a baseline of optimal performance.

RESULTS

The set of weights obtained by the GA for the urban warfare case base similarity function indicated that some attributes are much more important for assessing similarity than others. The 235 weights found were fairly evenly distributed over the full 0.0 – 10.0 range of possible weight values. Some of the most highly weighted features in this specific application, each having a weight above 9.8, were (our tactics = blitzing attack), (enemy abstract tactics = sewer battle), (arms = missiles), and (results = failed to expel invading force). Some of the lowest weighted features, each having a weight below 0.04, were (our tactics = aerial bombardment), (assessments = possible hasty attack), and (civilian actions = civilians serve as guides). Input features had an average weight of 4.43, while inferred attributes had an average weight of 4.17.

Our domain-independent approach to story interpretation appears, in limited testing to date, to work reasonably well. For example, the following excerpt from a made-up urban warfare story that is not in the case base,

We were a conventional army controlling a city in the middle of a war zone. The enemy militia was trying to take the city, and we had to defend it. We had virtually no re-supply lines and were poorly supplied. We had some small arms, some heavy machine guns, and some IEDs. In anticipation of the enemy advance we set up some fortified positions with heavy machine guns from which we could strafe the streets with fire. The public was not entirely supportive of the battle, so we wanted to try very hard to avoid civilian casualties. In the morning, three enemy tank columns entered the narrow streets of the city. We had to expel this invading force. From the rooftops, we began dropping grenades on to the enemy tanks. We detonated explosives that we had planted in buildings, exploding the buildings on to the approaching forces. The enemy attack choppers were ineffective in the battle as we kept them at bay with our SAMs….

is readily processed by the interpretation system. The abductive keyword parser processes this text and extracts the correct values for all of the input attributes, such as (arms = stingers), (arms = homemade explosives), (public opinion = public skeptical), and (tolerance for civilian casualties = low). Inferred attribute values include (inferred quality of intelligence = adequate), (quality of combined arms usage = poor), (assessments = enemy likely has outside support), and (suggested course of action = cut off enemy's outside support).

The similarity measure using the best set of weights derived by the GA identifies two episodes from a 1948 battle in the mandate of Palestine as being the most similar to the example episode above. Examining these two retrieved past cases, their similarity to the episode given above is readily apparent. In both cases an invading force of tanks is repelled by attacking them from above with explosives and by exploding buildings onto the tanks as they pass by. Examination of these stories by a human operator has the potential of leading to a number of new and useful inferences. For instance, in the second episode from the Palestine story, the enemy learns to deploy infantry support along with tanks, and this could be a useful point for a military commander. While our system currently does not have the ability to make inferences from retrieved cases, methods used in past case-based reasoning systems could be effective in this regard.

To assess in more general terms the effectiveness of the GA in deriving appropriate weights for application-specific case retrieval, we generated these weights using just the input features, just the inferred features, and both (10 trials with each), using the 30 cases in the urban warfare case base. The results of the similarity function's performance are given in Table 1. When the GA evolved weights using all of the stories in the case base, and then the rank assigned to the a priori human-specified best matching case was determined for each case in the case base using the similarity function, the mean rank was 1.37 if all features were used, 1.43 if just input features were used, and 2.87 if just inferred/output features were used. While using all features thus did marginally best, the contribution of the inferred features was almost negligible.

Table 1. Performance of similarity function using different sets of attribute.

	ALL STORIES			LEAVE ONE OUT		
ATTRIBUTES USED IN TRAINING	ALL	INPUT	OUTPUT	ALL	INPUT	OUTPUT
AVG. RANK GIVEN TO MOST SIMILAR	1.37	1.43	2.87	3.33	3.37	6.20
STANDARD DEVIATION OF RANK	0.72	0.73	2.26	2.51	2.76	6.58

To test the ability of this approach to generalize to new cases, we repeated the above study, but now using a leave-one-out strategy. In this situation, the mean rank was 3.33 if all features were used, 3.37 if just input features were used, and 6.20 if just inferred/output features were used. Though not as accurate as when all episodes are used, it suggests that one could use a strategy of retrieving three or four cases in general. The difference in performance between when all features were used and when just input features were used is not statistically significant. However, the difference between using just output features and either all features or just input features is statistically significant, at a higher than 95% confidence level.

DISCUSSION

In this project, we explored integrating reasoning from both general knowledge and from individual cases within a single framework. To this end, we investigated the feasibility of creating an application-independent approach to interpreting multi-episode "stories" that combine knowledge based reasoning methods with the retrieval of past most similar cases. In this paper we focused on whether the inferences made by a knowledge-based reasoning process can contribute to identifying the most relevant past cases during problem solving. Specifically, we explored the value of input features of a problem versus the value of inferred features in the retrieval of most relevant past cases.

For the urban warfare domain, we found that the input features of an episode are more important than the inferred features when attempting to assess the similarity of episodes. Our similarity measure performed significantly better when using only input attributes to assess similarity than when using only inferred attributes. This suggests that there is some information or relationships among the input attributes that our current knowledge, mostly in the form of production rules, simply does not capture. If this is correct, it supports the hypothesis that integrating knowledge based and case based reasoning is synergistic, because it suggests that the information in specific cases may be different from that inferred using general principles in an application domain. This was illustrated by the specific example case/story above where general rules deduced, for

example, that the enemy force probably had outside support and that a reasonable course of action would be to try to disrupt this support, while both retrieved cases included the sensible point that infantry should accompany tank incursions in an urban setting to help prevent attacks from above. In a limited sense, this can be viewed as a kind of "ensemble reasoning". Of course, it is entirely possible that a different set of inferred attributes might be more informative. The inferred attributes that we used were based on an a priori conception of useful inference without consideration of their utility for case retrieval. Our results are also limited to the specific domain of urban warfare, and it is unclear whether they will generalize to other areas like medical or legal reasoning.

The GA-derived weights used in the similarity function did not generalize extremely well. As expected, the ability to identify the a priori human-selected most similar case declined when a leave-one-out strategy was used to evaluate case retrievals. However, the results were still quite reasonable if one is willing to allow a system to retrieve a few apparently best cases rather than just the single best case. Using both input and output features to assess similarity of cases continued to result in the highest performance, but with inferred features being of negligible value.

An important direction for future work is the integration of additional reasoning methods with case-based reasoning. Some central questions are whether alternative reasoning methods or different inferred attributes can help improve case retrieval, how the results described here will generalize to domains other than urban warfare, and whether knowledge-driven inferences can contribute to more powerful case-based reasoning in general through better case adaptation/modification and storage. A more objective way to evaluate the performance of these tasks is also desirable. In particular, cost effective methods are needed for replacing the "gold standard" with more objective and precise ways of measuring relevance between cases. One possible approach would be the development of a simulated environment where agents could use case-based reasoning to solve problems. In such a system, the true relevance of past stories to a current problem could be determined by measuring the performance of the agent in addressing a new problem when recalling different past cases.

Acknowledgements: Supported by DARPA (F306029910552, FA87500520272), and by award 4400089699 from Raytheon Corporation.

REFERENCES

Antal J & Gericke B. *City Fights*, Ballantine, 2003.

Anon. http://www.wikipedia.org/wiki/Battle_of_Mogadishu,Wikipedia (25 Jan 2006).

Dubitzky W & Azuaje F. A genetic algorithm and growing cell structure approach to learning case retrieval structures, in *Soft Computing in Cased Based Reasoning*, S. Pal et al, eds), Springer-Verlag, 2001, 115-146

Grau L. *The Bear Went Over the Mountain*, National Defense University Press, 1996.

Herman D (ed.) *Narrative Theory and Cognitive Sciences*, CSLI, 2003.

Josephson J & Josephson S (eds.) *Abductive Inference*, Cambridge Univ. Press, 1994

Keegan J. *A History of Warfare*, Vintage, 1994.

Marling C, Sqalli M, Rissland E, Munoz-Avila H, & Aha D. Case-Based Reasoning Integrations, *AI Magazine,* 23, 2002, 69-86.

Oliker, Olga. *Russia's Chechen Wars 1994-2000*, RAND, 2001.

Pal S & Shiu S. *Foundations of Soft Case-Based Reasoning*, Wiley-Interscience, 2004.

Reggia, J. Abduction, *Encyclopedia of AI*, John Wiley, 1992, 2-3.