Knowledge Representation Considerations for a Domain Independent Semantic Parser

Mark Jones, Patrick Demasco, Kathleen McCoy, & Christopher Pennington

Applied Science and Engineering Laboratories
University of Delaware/ A.I. duPont Institute
Wilmington, Delaware USA

Abstract

We have previously presented the overall system design for the Compansion system and have discussed further enhancements to the parsing component that provides domain independent processing. In this paper, we discuss the knowledge representation scheme currently under development. The major improvements include a hierarchical-based lexicon representation, the use of case frame preferences for word role disambiguation, and improvements to the parsing logic that increase the overall system robustness.

Background

This work is part of an augmentative communication project being conducted at the Applied Science and Engineering Laboratories at the University of Delaware and the A.I. duPont Institute. The goal of this project is to increase the communication rate of physically disabled individuals via Natural Language Processing techniques.

We wish to take as input a compressed message (i.e., one containing mainly the content words of the desired utterance) from the disabled individual and generate a syntactically and semantically well-formed sentence. For a description of the Sentence Compansion system and the Semantic Parser's role in it see, (McCoy et al., 90). This paper builds on the previous work by describing recent insights into the knowledge representation scheme used by the semantic parser.

Statement of the Problem

The Semantic Parser (see also (Small &Rieger, 82)) is responsible for determining the semantic role being played by each input word. It must determine which word is the verb, what role each noun phrase plays with respect to the verb (e.g., actor, theme), and what modification relationships are present.

These inferences must be based on stored knowledge about individual words and possible word relationships. In our previous efforts, we utilized a a non-hierarchical word categorization scheme, and represented possible word relationships with relatively simple deterministic heuristics. While this approach proved satisfactory for relatively small vocabularies and simple sentence structures, it became necessary to consider substantial improvements to this aspect of the parser. In addition, we wanted to increase the robustness of the parser to accommodate ill-formed input.

Approach

Our approach is based on a Case Frame-based representation of sentence structure with word roles stored in hierarchical data structures. Heuristics employed to fill Case Roles represent uncertainty with preference scales. These preferences allow us to calculate a total confidence level for each potential parse. Finally enhancements to the parser logic allow us to infer the likely role of a word that is not explicitly stored in the lexicon.

<u>Case Frame Representation</u>

The output of the parser is in the form of Case Frames (Fillmore, 77). The main idea behind case frames is that in a sentence there is a fixed number of roles that objects can play with respect to the main verb. Given the input: [John break hammer], the parser will return the semantic parse below.

```
((43 DECL
    (VERB (LEX BREAK))
    (AGEXP (LEX JOHN))
    (THEME (LEX HAMMER))
    (TENSE PRES)))
```

This parse is consistent with the sentence, "John breaks the hammer." The first line gives a confidence value for the parse, and says that this is a declarative sentence. The second line states that the main verb of the sentence is *break*. The third line states that the AGEXP (doer) of the break action is *John*. The next line says that what is being broken is a *hammer.*

Given the input: [John tell Mary Joke Sue], the parser will return (among others) the semantic parse below.

```
(71 DECL
    (VERB (LEX TELL))
    (AGEXP (LEX JOHN))
    (THEME (LEX JOKE))
    (GOAL (LEX MARY))
    (BENEF (LEX SUE))
    (TENSE PRES))
```

This parse is consistent with the sentence, "John tells a joke to Mary for Sue." There are a few new cases that should be explained. Note that the THEME is *joke* and not *Mary*; *joke* is what is being told. *Mary* is the receiver, the GOAL. Finally, the act is done for *Sue*; she is the BENEFiciary.

Knowledge Representation

The knowledge used by the parser must be as great as possible because it cannot rely on syntactic information. Also, this knowledge must be domain independent. The parser utilizes several knowledge hierarchies of which two are particularly important. The object hierarchy captures generalizations about nouns. The verb hierarchy captures generalizations about verbs.The main verb of the sentence is key in predicting the semantic structure of the sentence. The layout of the verb hierarchy is motivated by work in systemic grammar (Halliday85). There are two general types of heuristics: Those that are semantic in nature, and those that are more idiosyncratic to the verb, more syntactic in nature.

Idiosyncratic Case Constraints

The idiosyncratic case constraints are called idiosyncratic because they are attached to individual verbs, rather than inherited. There are two key properties that can be associated with each verb. They are Mandatory and Forbidden. For example, the verb *hit* requires that the THEME be filled. The *mandatory* feature allows this to be represented. On the other hand, *hit* cannot accommodate a GOAL. This can also be represented in the system. This relates to traditional linguistics. Typically intransitive verbs forbid the filling of the THEME case: *die* cannot have a theme Words other than bi-transitives typically forbid filling the goal case: *give* can have a goal. The most common situation, that of the verb neither forbidding nor requiring a particular case, is represented by the absence of either feature.

Semantic Case Preferences

The semantic preferences differ from the syntactic predictions in a number of ways. First, these semantic preferences are not as definite, they are much fuzzier in nature, thus the term *preference*. These preferences are the basis for the heuristic values given to the output interpretations. Second, unlike the constraints, these semantic preferences are general enough to be inherited down a hierarchy of verbs. Third, these semantic preferences are closely tied to the object hierarchy. The case constraints have no interest in how the object knowledge base is structured.

Semantic preferences rely on a numeric scale ranging from 1 for low preference to 4 for high preference. In this scale, 1 and 4 are for special cases. 4 signifies that the binding is exceptionally appropriate. 1 signifies that the binding is only appropriate in special cases. For normal situations, the ratings of 2 and 3 are used. At this point, this granularity seems appropriate for our level of inferencing.

Case Importance Preference - This preference represents how important it is to fill a particular case in the frame. This is much more flexible than mandatory and forbidden which were described previously. For example, with material verbs such as *kick*, it seems much more likely that the role of THEME will be filled than that of BENEFiciary. To represent this, a higher value (3 on the 1 to 4 scale) is given as the preference of filling the THEME case, while a lower value (1) is given as the preference for filling the BENEFiciary case.

Case Filler Preference - This preference is directly related to the object hierarchy. Here, what kinds of objects should be playing the role is represented, this along with a preference of how reasonable such a binding seems. For example, the preference for filling the BENEFiciary case for most verbs is: ((human 3) (organizat 2) (animate 2)). This means that 3 points (again on the 1 to 4 scale) are given for binding a human in the given role. A binding of organizations, such as the A.C.L.U., or animate objects yield two points. This specification may seem ambiguous, is not any human also animate? True, the solution used in this system is that if a binding can achieve more than one score, the highest of the scores is used. Also note that a list such as the one just given is considered exclusive. In this example, it means that any verb following the stated pattern for BENEFiciaries will not allow objects that do not have as ancestors one of the three types given. A chair (inanimate object) would not be considered as a BENEFiciary.

The inheritance mechanism for the case importance and the case filler preferences is rather simple. Those preferences stated by the highest ancestors of the verb hold preferences that are reasonable in general. If conflicting information is given by more specific (lower) ancestor of the verb, the more specific information will be recognized.

Higher-Order Case Preferences - The mechanisms for Fill-Case and the Fill-Case-with-what preferences are limited in scope only to one role at a time (e.g., BENEFiciary). The Higher-Order Preferences fill the need for some more unifying heuristics With this power we can represent the following: If a non-human animate (e.g., dog) is the AGENT of a material process, it is quite unlikely that an instrument is being used.

Unknown Words

The power of this knowledge representation scheme iprovides robustness in parsing ability. The system is able to make some sense out of unknown words present in the input stream. If the parser knows the main verb of the sentence it can infer the role of the unknown word and the type of object that is represented. The parser assumes that an unknown word is an object. It then creates multiple senses of the unknown word; one sense for place, tool, food, etc. These senses are chosen to cover the range of objects yet not be too specific. Because multiple word senses are treated as mutually exclusive, it tries each sense separately. The heuristic ratings allow the interpretation(s) with the best word sense to rise to the top, and a moderately intelligent guess of the unknown word is achieved. The information inferred about the unknown word can be passed on to and referenced by the processes that follow the semantic parser. The table below lists several examples followed by the case that the unknown word (XXX) is interpreted to fill, and the type of object in the object hierarchy that the object is interpreted to be.

[John break window XXX]	INSTR tool
[John eat XXX fork]	THEME ingestible
[John eat pizza XXX]	INSTR tool
[John tell XXX Mary]	THEME abstract
[John go XXX]	LOC place
[XXX carry paper]	AGEXP animate and ergative-object (tie)

Implications

This approach takes advantage of several important generalizations. First, the object and verb hierarchies capture needed generalizations. Also, the preferences are distributed among the verbs in a motivated manner. This approach lends elegance to the system, and makes it easier to enhance.

Distinctions between the knowledge have been well placed. First, separating the idiosyncratic constraints (what roles must and must not be filled) from the preferences (what roles should be filled, and with what) is useful. This is key to the elegance of the knowledge hierarchies, because such behaviors cut across different dimensions. The notion of transitivity was used to explain the more syntactic knowledge. Such knowledge as transitivity clearly cuts along a different dimension than that of meaning. For example, the parser encodes the words eat and swallow as semantically equivalent. However, swallow cannot typically have an instrument, while eat may.

Previous approaches to representing such knowledge have not distinguished between that captured in case importance and case filler preferences. Without this distinction, statements such as those of case importance cannot be made. Recall, that for many material verbs, such as hit, filling the THEME role is very important. But without such information, the parser would mistakenly consider Mary in [John hit Mary] a BENEFiciary, because, other things being equal, people are highly correlated to the BENEFiciary role. Another major advantage of this system is its use of heuristics. Not only can the system handle [John break hammer], but also [John break hammer window], where hammer now plays a different role (Instrument). Through its robust heuristics, it can recognize the preferred interpretation of this message.

Discussion

Although the parser has been radically changed in the last year, it already captures the functionality of the previous system, including inferring agents and verbs in some situations.

With these theoretical improvements in the semantic parser, we come closer to our goal of making available an augmentative communication system which takes advantage of the power of research in the field of Natural Language Processing.

Acknowledgments

This work is supported by Grant Number H133E80015 from the National Institute on Disability and Rehabilitation Research. Additional support has been provided by the Nemours Foundation.

References

C.J. Fillmore.The case for case reopened. In P. Cole and J.M. Sadock, editors, *Syntax and Semantics VIII Grammatical Relations*, pages 59-81, Academic Press, New York, 1977.

Halliday M. A. K. *An Introduction to Functional Grammar*. Edward Arnold, London England, 1985

K. McCoy, P. Demasco, M. Jones, C. Pennington, and C. Rowe. A Domain Independent Semantic Parser For Companion. In *Proceedings of the 13th Annual RESNA conference*, pages 187-188, RESNA, Washington, DC., June 1990.

S. Small and C. Rieger. Parsing and comprehending-with word experts (a theory and its realization). In Wendy G. Lehnert and Martin H. Ringle, Editors, *Strategies for Natural Language Processing*, 1982.

Contact

Mark Jones
Applied Science and Engineering Laboratories
A.I. duPont Institute
P.O. Box 269
Wilmington, DE 19899
Email: jones@udel.edu