TOWARDS A SELF-EXTENDING PARSER

Jaime G. Carbonell
Department Of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

This paper discusses an approach to incremental learning in natural language processing. The technique of projecting and integrating semantic constraints to learn word definitions is analyzed as implemented in the POLITICS system. Extensions and improvements of this technique are developed. The problem of generalizing existing word meanings and understanding metaphorical uses of words is addressed in terms of semantic constraint integration.

1. Introduction

Natural language analysis, like most other subfields of Artificial Intelligence and Computational Linguistics, suffers from the fact that computer systems are unable to automatically better themselves. Automated learning is considered a very difficult problem, especially when applied to natural language understanding. Consequently, little effort has been focused on this problem. Some pioneering work in Artificial Intelligence, such as AM [1] and Winston's learning system [2] strove to learn or discover concept descriptions in well-defined domains. Although their efforts produced interesting ideas and techniques, these techniques do not fully extend to a domain as complex as natural language analysis.

Rather than attempting the formidable task of creating a language learning system, I will discuss techniques for incrementally increasing the abilities of a flexible language analyzer. There are many tasks that can be considered "incremental language learning". Initially the learning domain is restricted to learning the meaning of new words and generalizing existing word definitions. There are a number of A.I. techniques, and combinations of these techniques capable of exhibiting incremental learning behavior. I first discuss FOULUP and POLITICS, two programs that exhibit a limited capability for incremental word learning. Secondly, the technique of semantic constraint projection and integration, as implemented in POLITICS, is analyzed in some detail. Finally, I discuss the application of some general learning techniques to the problem of generalizing word definitions and understanding metaphors.

2. Learning From Script Expectations

Learning word definitions in semantically-rich contexts is perhaps one of the simpler tasks of incremental learning. Initially I confine my discussion to situations where the meaning of a word can be learned from the immediately surrounding context. Later I relax this criterion to see how global context and multiple examples can help to learn the meaning of unknown words.

The FOULUP program [3] learned the meaning of some unknown words in the context of applying a script to understand a story. Scripts [4, 5] are frame-like knowledge representations abstracting the important features and causal structure of mundane events. Scripts have general expectations of the actions and objects that will be encountered in processing a story. For instance, the restaurant script expects to see menus, waitresses, and customers ordering and eating food (at different pre-specified times in the story).

FOULUP took advantage of these script expectations to conclude that items referenced in the story, which were part of expected actions, were indeed names of objects that the script expected to see. These expectations were used to form definitions of new words. For instance, FOULUP induced the meaning of "Rabbit" in, "A Rabbit veered off the road and struck a tree," to be a self-propelled vehicle. The system used information about the automobile accident script to match the unknown word with the script-role "VEHICLE", because the script knows that the only objects that veer off roads to smash into road-side obstructions are self propelled vehicles.

3. Constraint Projection in POLITICS

The POLITICS system [6, 7] induces the meanings of unknown words by a one-pass syntactic and semantic constraint projection followed by conceptual enrichment from planning and world-knowledge inferences. Consider how POLITICS proceeds when it encounters the unknown word "MPLA" in analyzing the sentence:

"Russia sent massive arms shipments to the MPLA in Angola."

Since "MPLA" follows the article "the" it must be a noun, adjective or adverb. After the word "MPLA", the preposition "in" is encountered, thus terminating the current prepositional phrase begun with "to". Hence, since all well-formed prepositional phrases require a head noun, and the "to" phrase has no other noun, "MPLA" must be the head noun. Thus, by projecting the syntactic constraints necessary for the sentence to be well formed, one learns the syntactic category of an unknown word. It is not always *possible* to *narrow* the categorization of a word to a single syntactic category from one example. In such cases, I propose intersecting the sets of possible syntactic categories from more than one sample use of the unknown word until the intersection has a single element.

POLITICS learns the meaning of the unknown word by a similar, but substantially more complex, application of the same principle of projecting constraints from other parts of the sentence and subsequently integrating these constraints to construct a meaning representation. In the example above, POLITICS analyzes the verb "to send" as either an ATRANS or a PTRANS. (Schank [8] discusses the Conceptual Dependency *case frames*. Briefly, a PTRANS is a physical transfer of location, and an ATRANS is an abstract transfer of ownership, possession or control.) The reason why POLITICS cannot decide on the type of TRANSfer is that it does not know whether the destination of the transfer (i.e., the MPLA) is a location or an agent. Physical objects, such as weapons, are PTRANSed to locations but ATRANSed to agents. The conceptual analysis of the sentence, with MPLA as yet unresolved, is diagrammed below:

```
                                RECIP01 <is> LOC val *ANGOLA*
                                                    |
                                                    |
                                                  (MPLA)
               ATRANS    o                   d |--> RECIP01
*RUSSIA* <=>     or     <--- *WEAPON* <---|
               PTRANS              |            |--< *RUSSIA*
                                   |
                                   |
                         *WEAPON* <is> NUMBER val (> NORM)
```

What has the analyzer learned about "MPLA" as a result of formulating the CD case frame? Clearly the MPLA can only be an actor (i.e., a person, an institution or a political entity in the POLITICS domain) or a location. Anything else would violate the constraints for the recipient case in both ATRANS and PTRANS. Furthermore, the analyzer knows that the location of the MPLA is inside Angola. This item of information is integrated with the case constraints to form a partial definition of "MPLA". Unfortunately both locations and actors can be located inside countries; thus, the identity of the MPLA is still not uniquely resolved. POLITICS assigns the name RECIP01 to the partial definition of "MPLA" and proceeds to apply its inference rules to understand the political implications of the event. Here I discuss only the inferences relevant for further specifying the meaning of "MPLA".

4. Uncertain Inference in Learning

POLITICS is a goal-driven inferencer. It must explain all actions in terms of the goals of the actors and recipients. The emphasis on inducing the goals of actors and relating their actions to means of achieving these goals is integral to the theory of subjective understanding embodied in POLITICS. (See [7] for a detailed discussion.) Thus, POLITICS tries to determine how the action of sending weapons can be related to the goals of the Soviet Union or any other possible actors involved in the situation. POLITICS knows that Angola was in a state of civil war; that is, a state where political factions were exercising their goals of taking military and, therefore, political control of a country. Since possessing weapons is a precondition to military actions, POLITICS infers that the recipient of the weapons may have been one of the political factions. (Weapons are a means to fulfilling the goal of a political faction, therefore POLITICS is able to explain why the faction wants to receive weapons.) Thus, MPLA is inferred to be a political faction. This inference is integrated with the existing partial definition and found to be consistent. Finally, the original action is refined to be an ATRANS, as transfer of possession of the weapons (not merely their location) helps the political faction to achieve its military goal.

Next, POLITICS tries to determine how sending weapons to a military faction can further the goals of the Soviet Union. Communist countries have the goal of spreading their ideology. POLITICS concludes that this goal can be fulfilled only if the government of Angola becomes communist. Military aid to a political faction has the standard goal of military takeover of the government. Putting these two facts together, POLITICS concludes that the Russian goal can be fulfilled if the MPLA, which may become the new Angolan government, is Communist. The definition formed for MPLA is as follows:

```
Dictionary entry:
(DPS MPLA (POS  NOUN (TYPE PROPER)))
          (TOK *MPLA*))

Memory entry:
(DPS *MPLA* (ISA . *FACTION*)
            (PARTOF . *ANGOLA*)
            (IDEOLOGY . *COMMUNIST*)
            (GOALS: ((ACTOR (*MPLA*) IS
                        (SCONT OBJECT (*ANGOLA*)
                              VAL (10))))))
```

The reason why memory entries are distinct from dictionary definitions is that there is no one-to-one mapping between the two. For instance, "Russia" and "Soviet Union" are two separate dictionary entries that refer to the same concept in memory. Similarly, the concept of SCONT (social or political control) abstracts information useful for the goal-driven inferences, but has no corresponding entry in the lexicon, as I found no example where such concept was explicitly mentioned in newspaper headlines of political conflicts (i.e., POLITICS' domain).

Some of the inferences that POLITICS made are much more prone to error than others. More specifically, the syntactic constraint projections and the CD case-frame projections are quite certain, but the goal-driven inferences are only reasonable guesses. For instance, the MPLA could have been a plateau where Russia deposited its weapons for later delivery.

5. A Strategy for Dealing with Uncertainty

Given such possibilities for error, two possible strategies to deal with the problem of uncertain inference come to mind. First, the system could be restricted to making only the more certain constraint projection and integration inferences. This does not usually produce a complete definition, but the process may be iterated for other exemplars where the unknown word is used in different semantic contexts. Each time the new word is encountered, the semantic constraints are integrated with the previous partial definition until a complete definition is formulated. The problem with this process is that it may require a substantial number of iterations to converge upon a meaning representation, and when it eventually does, this representation will not be as rich as the representation resulting from the less certain goal-driven inferences. For instance, it would be impossible to conclude that the MPLA was Communist and wanted to take over Angola only by projecting semantic constraints.

The second method is based on the system's ability to recover from inaccurate inferences. This is the method I implemented in POLITICS. The first step requires the detection of contradictions between the inferred information and new incoming information. The next step is to assign blame to the appropriate culprit, i.e., the inference rule that asserted the incorrect conclusion. Subsequently, the system must delete the inaccurate assertion and later inferences that depended upon it. (See [9] for a model of truth maintenance.) The final step is to use the new information to correct the memory entry. The optimal system within my paradigm would use a combination of both strategies - it would use its maximal inference capability, recover when inconsistencies arise, and iterate over many exemplars to refine and confirm the meaning of the new word. The first two criteria are present in the POLITICS implementation, but the system stops building a new definition after processing a single exemplar unless it detects a contradiction.

Let us briefly trace through an example where POLITICS is told that the MPLA is indeed a plateau after it inferred the meaning to be a political faction.

```
[ POLITICS run -- 2/06/78 ]

*(INTERPRET US-CONSERVATIVE)
INPUT STORY:  Russia sent massive arms shipments
              to the MPLA in Angola.

PARSING...(UNKNOWN WORD: MPLA)
(SYNTACTIC EXPECTATION: NOUN)
(SEMANTIC EXPECTATION: (FRAME: (ATRANS PTRANS) SLOT: RECIP
                        REQ: (LOC ACTOR))) COMPLETED.

CREATING NEW MEMORY ENTRY: *MPLA*
INFERENCE:  *MPLA* MAY BE A POLITICAL FACTION OF *ANGOLA*
INFERENCE:  *RUSSIA* ATRANS *ARMS* TO *MPLA*
INFERENCE:  *MPLA* IS PROBABLY *COMMUNIST*
INFERENCE:  GOAL OF *MPLA* IS TO TAKE OVER *ANGOLA*
INSTANTIATING SCRIPT:  $AIDMF
INFERENCE:  GOAL OF *RUSSIA* IS *ANGOLA* TO BE *COMMUNIST*

[ Question-answer dialog ]
*What does the MPLA want the arms for?
 THE MPLA WANTS TO TAKE OVER ANGOLA USING THE WEAPONS.

What might the other factions in Angola do?
 THE OTHER FACTIONS MAY ASK SOME OTHER COUNTRY FOR ARMS.

[ Reading further input ]
 INPUT STORY:   *The Zungabi faction operating from the MPLA
                 plateau received the Soviet weapons.

PARSING...COMPLETED.

CREATING NEW MEMORY ENTRY: *ZUNGABI*
ACTIVE CONTEXT APPLICABLE: $AIDMF

C1  ISA CONFLICT:  *MPLA* ISA (*FACTION* *PLATEAU*)
(ACTIVATE (INFCHECK C1)) REQUESTED
C2  SCRIPT ROLE CONFLICT:
    (&AID-RECIP IN $AIDMF) = *MPLA* AND *ZUNGABI*
(ACTIVATE (INFCHECK C2)) REQUESTED

(INFCHECK C1 C2) INVOKED:
 ATTEMPT TO MERGE MEMORY ENTRIES: (*MPLA* *ZUNGABI*)...FAILURE
 INFERENCE RULE CHECKED (RULE#F1 . $AIDMF)...OK
 INFERENCE RULE CHECKED (RULE#G8)...CONFLICT!
 DELETING RESULT OF RULE#G8

 C1 RESOLVED: *MPLA* ISA *PLATEAU* IN *ANGOLA*
 C2 RESOLVED: (&AID-RECIP IN $AIDMF) = *ZUNGABI*

 REDEFINING *MPLA* AS *ZUNGABI*...COMPLETED.
 CREATING NEW *MPLA* MEMORY ENTRY...COMPLETED.
```

POLITICS realizes that there is an inconsistency in its interpretation when it tries to integrate "the MPLA plateau" with its previous definition of "MPLA". Political factions and plateaus are different conceptual classes. Furthermore, the new input states that the Zungabi received the weapons, not the MPLA. Assuming that the input its correct, POLITICS searches for an inference rule to assign blame for the present contradiction. This is done simply by temporarily deleting the result of each inference rule that was activated in the original interpretation until the contradiction no longer exists. The rule that concluded that the MPLA was a political faction is found to resolve both contradictions if deleted. Since recipients of military aid must be political entities, the MPLA being a geographical location no longer qualifies as a military aid recipient.

Finally, POLITICS must check whether the inference rules that depended upon the result of the deleted rule are no longer applicable. Rules, such as the one that concluded that the political faction was communist, depended upon there being a political faction receiving military aid from Russia. The Zungabi now fulfills this role; therefore, the inferences about the MPLA are transfered to the Zungabi, and the MPLA is redefined to be a plateau. (Note: the word "Zungabi" was constructed for this example. The MPLA is the present ruling body of Angola.)

6. Extending the Project and Integrate Method

The POLITICS implementation of the project-and-integrate technique is by no means complete. POLITICS can only induce the meaning of concrete or proper nouns when there is *sufficient* contextual information in a single exemplar. Furthermore, POLITICS assumes that each unknown word will have only one meaning. In general it is useful to realize when a word is used to mean something other than its definition, and subsequently formulate an alternative definition.

I illustrate the case where many examples are required to narrow down the meaning of a word with the following example: "Johnny told Mary that if she didn't give him the toy, he would <unknown-word> her." One can induce that the unknown word is a verb, but its meaning can only be guessed at, in general terms, to be something unfavorable to Mary. For instance, the unknown word could mean "take the object from", or "cause injury to". One needs more than one example of the unknown word used to mean the same thing in different contexts. Then one has a much richer, combined context from which the meaning can be projected with greater precision.

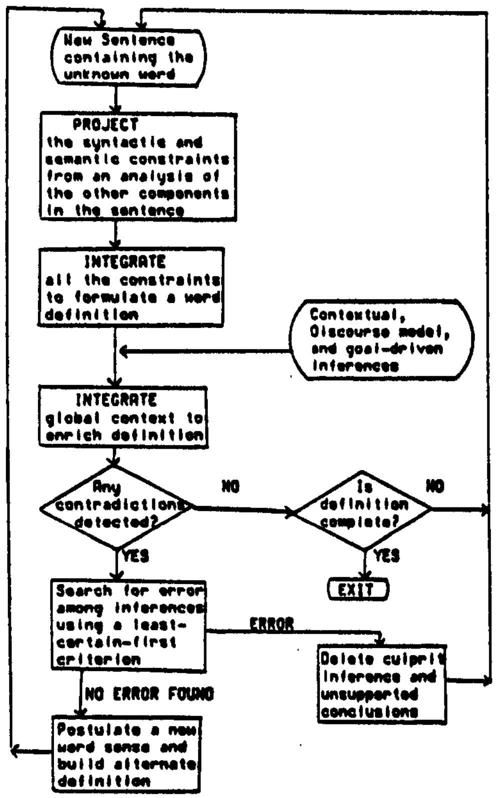
Figure 1 diagrams the general project-and-integrate algorithm. This extended version of POLITICS' word-learning technique addresses the problems of iterating over many examples, multiple word definitions, and does not restrict its input to certain classes of nouns.

7. Generalizing Word Definitions.

Words can have many senses, some more general than others. Let us look at the problem of generalizing the semantic definition of a word. Consider the case where "barrier" is defined to be a physical object that disenables a transfer of location. (e.g. "The barrier on the road is blocking my way.") Now, let us interpret the sentence, "Import quotas *form a barrier* to international trade." Clearly, an import quota is not a physical object. Thus, one can minimally generalize "barrier" to mean "anything that disenables a physical transfer of location."

Let us substitute "tariff" for "quota" in our example. This suggests that our meaning for "barrier" is insufficiently general. A tariff cannot disenable physical transfer; tariffs disenable willingness to buy or sell goods. Thus, one can further generalize the meaning of barrier to be: "anything that disenables any type of transfer". Yet, some trace of the

Figure 1: The project-and-integrate method for inducing new word and concept definitions



generalization process must be remembered because the original meaning is often preferred, or metaphorically referenced. Consider: "The trade barriers were lifted." and "The new legislation bulldozed existing trade barriers." These sentences can only be understood metaphorically. That is, one needs to refer to the original meaning of 'barrier" as a physical object, in order for "lifting" or 'bulldozing" to make sense. After understanding the literal meaning of a "bulldozed barrier", the next step is to infer he consequence of such an action, namely, the barrier no onger exists. Finally, one can refer to the generalized meaning of "barrier" to interpret the proposition that "The ew legislation caused the trade barriers to be no longer in xistence."

propose the following rules to generalize word definitions nd understand metaphorical references to their original, ormal definition:

1) If the definition of a word violates the semantic constraints projected from an interpretation of the rest of the sentence, create a new word-sense definition that copies the old definition minimally relaxing (i.e., generalizing) the violated constraint.

2) In interpreting new sentences always prefer the most specific definition if applicable.

3) If the generalized definition is encountered again in interpreting text, make it part of the permanent dictionary.

4) If a word definition requires further generalization, choose the existing most general definition and minimally relax its violated semantic constraints until a new, yet more general definition is formed.

5) If the case frame formulated in interpreting a sentence projects more specific semantic constraints onto the word meaning than those consistent with the entire sentence, interpret the word using the most specific definition consistent with the case frame. If the resultant meaning of the case frame is inconsistent with the interpretation of the whole sentence, infer the most likely consequence of the partially-build Conceptual Dependency case frame, and use this consequence in interpreting the rest of the sentence.

The process described by rule 5 enables one to interpret the metaphorical uses of words like "lifted" and "bulldozed" in our earlier examples. The literal meaning of each word is applied to the object case, (i.e., "barrier"), and the inferred consequence (i.e., destruction of the barrier) is used to interpret the full sentence.

8. Concluding Remarks

There are a multitude of ways to incrementally improve the language understanding capabilities of a system. In this paper I discussed in some detail the process of learning new words. In lesser detail I presented some ideas on how to generalize word meanings and interpret metaphorical uses of individual words. There are many more aspects to learning language and understanding metaphors that I have not touched upon. For instance, many metaphors transcend individual words and phrases. Their interpretation may require detailed cultural knowledge [10].

In order to place some perspective on project-and-integrate learning method, consider three general learning mechanisms capable of implementing different aspects of incremental language learning.

Learning by example. This is perhaps the most general learning strategy. From several exemplars, one can intersect the common concept by, if necessary, minimally generalizing the meaning of the known part of each example until a common subpart is found by intersection. This common subpart is likely to be the meaning of the unknown section of each exemplar.

Learning by near-miss analysis. Winston [2] takes full advantage of this technique. It may be usefully applied to a natural language system that can interactively generate utterances using the words it learned, and later be told whether it used those words correctly, whether it erred seriously, or whether it came close but failed to understand a subtle nuance in meaning.

Learning by contextual expectation. Essentially FOULUP and POLITICS use the method of projecting contextual expectations to the linguistic element whose meaning is to be induced. Much more mileage can be gotten from this method, especially if one uses strong syntactic constraints and expectations from other knowledge sources, such as a discourse model, a narrative model, knowledge about who is providing the information, and why the information is being provided.

9. References

1. Lenat, D. *AM: Discovery in Mathematics as Heuristic Search.* Ph.D. Th., Stanford University, 1977.

2. Winston, P. *Learning Structural Descriptions from Examples.* Ph.D. Th., MIT, 1970.

3. Granger, R. *FOUL-UP: A Program that Figures Out Meanings of Words from Context.* IJCAI-77, 1977.

4. Schank, R. C. and Abelson, R. P. *Scripts, Goals, Plans and Understanding.* Hillside, NJ: Lawrence Erlbaum, 1977.

5. Cullingford, R. *Script Application: Computer Understanding of Newspaper Stories.* Ph.D. Th., Yale University, 1977.

6. Carbonell, J. G. POLITICS: Automated Ideological Reasoning. *Cognitive Science 2*, 1 (1978), 27-51.

7. Carbonell, J. G. *Subjective Understanding: Computer Models of Belief Systems..* Ph.D. Th., Yale University, 1979.

8. Schank, R. C. *Conceptual Information Processing.* Amsterdam: North-Holland, 1975.

9. Doyle, J. Truth Maintenance Systems for Problem Solving. Master Th., M.I.T., 1978.

10. Lakoff, G. and Johnson, M. Towards an Experimentalist Philosopher: The Case From Literal Metaphor. In preparation for publication, 1979.