The Acquisition of Conceptual Structure for the Lexicon

James Pustejovsky
Department of Computer Science
Brandeis University
Waltham, MA 02254
617-736-2709
jamesp@brandeis.csnet-relay

Sabine Bergler
Computer and Information Science
University of Massachusetts
Amherst, MA. 01003
bergler@umass.csnet-relay

Abstract

There has recently been a great deal of interest in the structure of the lexicon for natural language understanding and generation. One of the major problems encountered has been the optimal organization of the enormous amounts of lexical knowledge necessary for robust NLP systems. Modifying machine readable dictionaries into semantically organized networks, therefore, has become a major research interest. In this paper we propose a representation language for lexical information in dictionaries, and describe an interactive learning approach to this problem, making use of extensive knowledge of the domain being learned. We compare our model to existing systems designed for automatic classification of lexical knowledge.

1. Introduction

In this paper we describe an interactive machine learning approach to the problem of making machine readable dictionaries useful to natural language processing systems. This is accomplished in part by making extensive use of the knowledge of the domain being learned. The domain model we are assuming is the ***Extended Aspect Calculus***, [Pustejovsky, 1987], where possible word (verb) meanings are constrained by how arguments may bind to semantic types. In the case of lexical meanings for words, if the semantic theory constrains what a possible word meaning can be, then the learning task is greatly simplified since the model specializes the most general rule descriptions. The system generates hypothesis instances for word meanings based on the domain model, for which the interactive user acts as credit assigner. Generalization proceeds according to the paths established by the model. We compare our framework to existing systems designed for automatic classification of lexical knowledge.

There are three points we wish to make in this paper:

- The semantic relations and connections between lexical items in a dictionary can be easily learned if a semantic model of the domain is used to bias the acquisition process.
- A theory of lexical semantics can act as this model, constraining what a possible word type is, just as a grammar constrains what an acceptable sentence is.
- An interactive knowledge acquisition device can improve the performance of purely algorithmic approaches to lexical hierarchy-formation.

The paper will be organized as follows. In the second section we discuss the lexical information necessary for robust natural language processing systems. In section three we outline a framework for encoding the lexical semantics associated with a word, the Extended Aspect Calculus. Then in section four we describe how to set up a dictionary environment for efficient lexical acquisition. Section five runs through the knowledge acquisition system, TULLY, which learns the semantic structure of verbs with the help of an interactive critic. Finally, we discuss how our system compares to previous attempts at lexical acquisition, and discuss directions for future research.

2. What is Useful Lexical Information for NLP?

One of the central issues currently being addressed in natural language processing is: what information is needed in the lexicon for a system, in order to perform robust analysis and generation [Cf. Ingria, 1986, Cumming, 1986]? We examine this issue in detail here, and review what seems to be the minimum requirements for any lexicon.

Let us begin with one of the most central needs for analysis and parsing of almost any variety: knowing the ***polyadicity*** of a relation; that is, how many arguments a verb or predicate takes. Although this would appear to be a straightforward problem to solve, there is still very little agreement on how to specify what is and isn't an argument to a relation. For example, the verb ***butter*** can appear with two, three, four, or apparently five arguments, as illustrated below.

(1) a. John buttered the toast.
b. John buttered the toast with a knife.
c. John buttered the toast with a knife in the kitchen.
d. John buttered the toast with a knife in the kitchen on Tuesday.

Some indication must be given, either explicitly or implicitly, of how many NPs to expect for each verb form. Ignoring how each argument is interpreted for now, we could represent ***butter*** as $butter(x,y)$, $butter(x,y,z)$, $butter(x,y,z,w)$, or $butter(x,y,z,w,v)$. Generally, we can make a distinction between the real arguments and the modifiers of a predicate.

Even with a clear method of determining what is a modifier, another problem is posed by verbs such as ***open***, ***melt***, ***sink***, and ***close***, called causative/inchoative pairs, and discussed in [Atkin et al, 1986]. These verbs typically have both an intransitive, noncausative reading (2b), and a transitive, causative reading (2a).

(2) a. Susan opened the door.
b. The door opened.

The issue here is whether there should be two separate entries for verbs like *open* – $open(x,y)$ and $open(x)$– or one entry with a rule relating the two forms – $open(x,y) \Leftrightarrow open(y)$.

The arguments to nominal forms seem even more variable than those for verbs. In fact, they are in general entirely optional. For example, the nominal ***destruction*** can appear by itself (3a), with one argument (3b), or with two arguments (3c).

(3) a. The destruction was widespread.
b. The destruction of the city took place on Friday.
c. The army's destruction of the city took place on Friday.

We will not consider nominal forms in this paper, however.

Knowing the number of arguments for a verb is obviously of no use if the lexicon gives no indication of where each one appears in the syntax. For example, in the simple active form of *butter*, the argument occuppying the subject slot is always going to be the x in the argument list, while the passive form changes this. For verbs such as *open*, however, arguments which perform different functions can occupy the subject position using the same form of the verb. We will term this the ***external argument*** specification, which must somehow be given by a word entry. The other side of this is knowling how the ***complements*** of a verb are syntactically realized; this is termed the ***subcategorization problem***. For example, ***give*** has two complement types, *NP NP* and *NP PP*, as shown in (4a) and (4b). That is, in one case, the VP contains an NP followed by an NP, and in the other, an NP followed by a PP.

(4) a. John gave Mary the book.
b. John gave the book to Mary.

In addition to these specifications, for some arguments it will be necessary to indicate certain "selectional properties". For example, the verb ***put*** as used in (5),

(5) Mary put the book on the shelf.

requires that the third argument be realized as a PP, and furthermore that this preposition be a locative ([+*Loc*]). Likewise, many verbs of transfer, such as ***give***, ***send***, and ***present***, require the indirect object to be marked with the preposition ***to*** ([+*to*]) if it follows the direct object in the syntax (cf. (4b)). This information must be associated with the verb somehow, presumably by a lexical specification of some sort.

What we have discussed so far is only the simplest syntactic information about a verb. The real difficulty comes when we try to give an account of the semantics for an entry. This is typically achieved in natural language processing systems by associating the arguments with "named relations", such as ***Agent***, ***Patient***, ***Instrument***, ***Actor***, etc. These are represented as case roles or thematic roles.[1]

With this additional information, the lexical entry for ***give***, for example, will now look something like (5), ignoring the finer details.

(5) $give(x,y,z)$: $x = Agent$, $y = Patient$, $z = Goal$, $x = External$. $if\ x = External\ then\ z \doteq [+to]$.

The information we have assembled thus far will still not be rich enough for the deep understanding or fluent generation of texts. For each lexical item, there are associated inferences which must be made, and those that can be made about a certain state of affairs. One class of inferences deals with the ***aspectual*** properties associated with a verb. This identifies a verb with a particular event-type, such as ***state***, ***process***, or ***event***. For example, from (6a), we can infer (6b), while from (7a), no such inference is possible (cf. (7b)).

(6) a. John is running.
b. $\models$ John has run.
(7) a. John is drawing a circle.
b. $\not\models$ John has drawn a circle.

What is at work here is the fact that the meanings of certain verbs seem to entail a termination or end point, while for other verbs this is not the case. Thus, "drawing a circle" and "building a house" are events which have logical culminations, while simply "running" or "walking" do not. These types of inferences interact crucially with tense information for the analysis of larger texts and discourses. For more detail see [Pustejovsky, 1987b].

Finally, to make lexicon entries useful for performing inferences about classes and categories, it is important to know how each entry fits into a semantic hierarchy or network. Cf. [Amsler, 1980, Touretzky, 1986].

Let us now review what requirements we have placed on the lexical specification of word entries (where we limit ourselves here to verbal forms).

The lexicon should specify:

1. How many arguments the verb takes (the polyadicity).
2. An indication of where each argument appears in the syntax:
   i. Which is the *external* argument; and
   ii. What the subcategorization(s) are.
3. Optionality or obligatoriness of arguments.
4. Selectional properties of the verb on its arguments;- i.e. what preposition types they must occur with, in addition to semantic features such as ***animacy***, ***count***, ***mass***, etc.
5. The case roles of the arguments, e.g. ***Agent***, ***Instrument***, etc.
6. The aspectual type of a verb; i.e. whether it is a ***state***, ***process***, or ***event***.
7. Categorization or type information; e.g. as expressed in a semantic hierarchy.

Having reviewed the basic needs for a NLP lexicon, we will now outline a representation framework for encoding this information.

3. A Model of Lexical Semantics

In this section we outline the semantic framework which defines our domain for lexical acquisition. The model we have in mind acts to constrain the space of possible word meanings, by restricting the form that lexical decomposition can take. In this sense it is similar to Dowty's theory of lexical decomposition ([Dowty, 1979]), but differs in some important respects.[2]

Lexical decomposition is a technique for assigning meanings to words in order to perform inferences between them. Generative semantics [Lakoff, 1972]] took this technique to its limit in determining word semantics, but failed, however, to provide an adequate theory of meaning. In the AI literature, primitives have been suggested and employed with varying degrees of success, [Schank, 1975, Wilks, 1975], but do tend to prove useful.

[1] We will use these terms interchangeably, although there are strictly speaking, technical distinctions made by many people. For further discussion of case roles, see Gruber's original work on the matter, [Fillmore, 1968], as well as Gruber's treatment of thematic relations, [Gruber, 1965], and as extended by [Jackendoff, 1972].

[2] Space does not permit us to compare the framework here with that of [Dowty , 1979] and [Hinrichs, 1985]. See [Pustejovsky, 1987b] for a full discussion.

The model we present here, the Extended Aspect Calculus, is a partial decomposition model, indicating only a subset of the total interpretation of a lexical item. Yet, as we see in the next section, this partial specification is very useful for helping structure dictionary entries. For a more detailed description of our model, see [Pustejovsky, 1987b].

In the current linguistic literature on case roles or thematic relations, there is little discussion on what logical connection exists between one θ-role and another. The most that is claimed is that there is a repertoire of thematic relations, *Agent*, *Theme*, *Patient*, *Goal*, *Source*, *Instrument*, *Locative*, *Benefactive*, and that every NP must carry at least one role. It should be remembered, however, that thematic relations were originally conceived in terms of the argument positions of semantic predicates such as CAUSE and DO, present in the decomposition of verbs. [3]

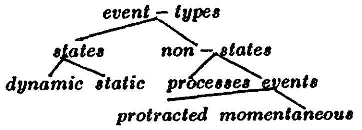
For example, the causer of an event (following [Jackendoff, 1976]) is defined as an *Agent* $CAUSE(x,e) \rightarrow Agent(x)$.

Similarly, the first argument position of the predicate GO is interpreted as Theme, as in $GO(x,y,z)$. The second argument here is the *SOURCE* and the third is called the *GOAL*.

Our model is a first-order logic that employs special symbols acting as operators over the standard logical vocabulary. These are taken from three distinct semantic fields. They are: ***causal***, ***spatial***, and ***aspectual***. The predicates associated with the causal field are: $Causer(C_1)$, $Causee(C_2)$, and $Instrument(I)$. The spatial field has two predicate types: *Locative* and *Theme*. Finally, the aspectual field has three predicates, representing three temporal intervals: t_1, beginning, t_2, middle, and t_3, end. From the interaction of these predicates all thematic types can be derived.[4]

Let us illustrate the workings of the calculus with a few examples. For each lexical item, we specify information relating to the argument structure and mappings that exist to each semantic field; we term this information the ***Thematic Mapping Index (TMI)***.[5]

Part of the semantic information specified lexically will include some classification into one of the following event-types (cf. [Kenny 1963], [Vendler 1967], [Ryle 1949], [Dowty 1979], [Bach, 1986]).



For example, the distinction between state, activity (or process), and accomplishment can be captured in the following way. A state can be thought of as reference to an unbounded interval, which we will simply call t_2; that is, the state spans this interval.[6] An activity or process can be thought of as referring to a designated initial point and the ensuing process; in other words, the situation spans the two intervals t_1 and t_2. Finally, an event can be viewed as referring to both an activity and a designated terminating interval; that is, the event spans all three intervals, t_1, t_2, and t_3.

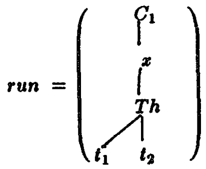
We assume that part of the lexical information specified for a predicate in the dictionary is a classification into some event-type as well as the number and type of arguments it takes. For example, consider the verb ***run*** in sentence (8), and ***give*** in sentence (9).

(8) John ran yesterday.

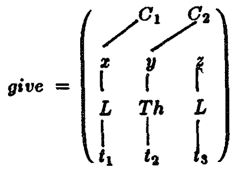
(9) John gave the book to Mary.

We associate with the verb ***run*** an aspect structure P (for process) and an argument structure of simply $run(x)$. For ***give*** we associate the aspect structure A (for accomplishment), and the argument structure $give(x,y,z)$. The Thematic Mapping Index for each is given below in (10) and (11).

(10)



(11)



The sentence in (8) represents a process with no logical culmination, and the one argument is linked to the named case role, *Theme*. The entire process is associated with both the initial interval t_1 and the middle interval t_2. The argument x is linked to C_1 as well, indicating that it is an *Actor* as well as a moving object (i.e. *Theme*). This represents one TMI for an activity verb.

The structure in (9) specifies that the meaning of ***give*** carries with it the supposition that there is a logical culmination to the process of giving. This is captured by reference to the final subinterval, t_3. The linking between x and the L associated with t_1 is interpreted as *Source*, while the other linked arguments, y and z are *Theme* (the book) and *Goal*, respectively. Furthermore, x is specified as a *Causer* and the object which is marked *Theme* is also an affected object (i.e. *Patient*). This will be one of the TMIs for an accomplishment.

In this fashion, we will encode the thematic and aspectual information about lexical items. This will prove to be a useful representation, as it allows for hierarchical organization among the indexes and will be central to our learning algorithm and the particular way specialization and generalization is performed on conceptual units. Essentially, the indexes define nodes in a tangled specialization tree, where the more explicitly defined the associations for a concept are, the lower in the hierarchy it will be.[7]

[3] Cf. [Jackendoff 1972, 1976] for a detailed elaboration of this theory.

[4] The presentation of the theory is simplified here, as we do not have the space for a complete discussion. See [Pustejovsky, 1987b] for discussion.

[5] [Marcus, 1987] suggests that the lexicon have some structure similar to what we are proposing. He states that a lexicon for generation or parsing should have the basic thematic information available to it.

[6] This is a simplication of our model, but for our purposes the difference is moot. A state is actually interpreted as a primitive homogeneous event-sequence, with downward closure. Cf. [Pustejovsky, 1987b],

[7] [Miller, 1985] argues that something like this is psychologically plausible, as well.

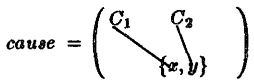
The most general concept types will be those indexes with a single link to one argument. [8]

4. Setting up the Environment

Before we describe the knowledge acquisition algorithm, we must define how to build the environment necessary for acquiring lexical information for a particular dictionary [Walker, 1986, Calzolari, 1984, Amsler 1984]. Although the specifics of the environment-setting will vary from dictionary to dictionary and from task to task, we are able to give a set of parameterizable features which can be used in all cases.

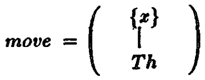
For each dictionary and task, the set of semantic primitives must be selected and specified by hand. These include all the entries corresponding to the operators from section 3, including *move*, *cause*, *become*, *be*, as well as aspectual operators such as *begin*, *start*, *stop*, etc.

For each primitive, we associate a thematic representation in terms of our model structure. For example, the particular word(s) in the dictionary that will refer to *cause* will have as their interpretation in the model, the following *partial thematic mapping index*:



This says that if cause is part of the definition of some term, we can assume that there are at least two argument places in that verb, and that one represents the causer, and the other the causee.

As another example, consider an entry with the primitive *move* in its definition. We can assume, in this case, that there is some argument which will associate with the *Theme* role.



Similar structures, what we term *partial TMIs*, can be defined for each primitive in the representation language.[9]

In addition to associating specific words in the dictionary with primitives in the representation language, we need to define the criteria by which we complete the association between arguments and thematic types. This is tantamount to claiming that case (or thematic) roles are simply sets of inferences associated with an argument of a verb. [10] For example, suppose we want to know whether the first argument of some verb should be assigned the role of *Agent* or *Instrument* of causation. We need to determine whether the argument is *animate* and *directly* the cause. These features will be assigned in the interactive session of the acquistion phase, and the associated role(s) will be assigned. Similar tests are used to determine whether something is a moving object, the source of an action, etc. Aspectual tests will determine whether something is an activity or accomplishment, etc. These tests are heuristics and can be modified as necessary by the user for his purposes.

Finally, another parameterizable component of the system deals with extracting specialized information from the dictionary entry. These are called *Thematic Specialists*. For example, consider a definition that defines some motion and contains the phrase *with the arm*. This is an example where additional information specifies *thematically incorporated* roles ([Gruber, 1968]). In other words, the instrument of the action is restricted to the arm. Another example would be the locative phrase *on water*, specifying the restriction on the location of the action. Each thematic specialist looks for particular patterns in the definition and proposes the associated thematic linking to the user in the interactive session.

5. The Learning Algorithm

Now that we have described how to set up the dictionary environment, we turn to the acquisition algorithm itself. We will illustrate each step in the procedure with an example run.[11]

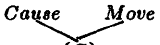
(1) *Select a Primitive:* Choose a primitive, p from the set of primitives, P, in the representation language. We begin with the intersective primitive *cause* and *move*.[12]

That is, we are narrowing in our learning mechanism to words whose entries contain both these primitives.

(2) *Form Candidate-Set:* Modified Head-Finding Algorithm ([Chodorow et al, 1985]). This returns a set of words, C with the primitives in their definitions, namely:

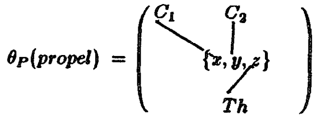
$$cause \cup move = \{turn, shake, propel, walk\}$$

(3) Place this set C into a tangled hierarchy, dominated by both *Cause* and *Move*.



(4) *Select a single candidate:* From C pick a candidate, c. We select *propel*.

(5) *Assign Partial Thematic Mapping Index:* This derives the partial TMI, termed $\theta_P(c)$.



(6) *Interactive Credit Assignment:* The TMI is completed interactively, where the user acts as credit assigner. Questions include: *Is x animate?*, *Does y move?* *Can x propel y for an hour?*, etc. The new information includes the aspectual class, selectional information, and of course the complete thematic mapping. The system conclude that x is the first

[9] One problem that we recognize, but have not addressed here, is multiple word sense. [Amsler, 1980] makes the point quite clear, that to fully disambiguate each entry and the entries defined in terms of them, is a major task.

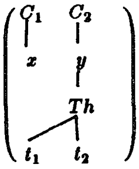
[10] This is hinted at in Miller and Johnson-Laird's 1976 pioneering work, and has recently been suggested, in a somewhat different form, by Dowty and Ladusaw (1986).

[11] We have based our environment on hand-coded definitions from the American Heritage Dictionary, [AHD, 1983]. Throughout the example, we have simplified or shortened the output.

[12] There is good reason to begin the search with entries containing two primitive terms. First, this will generally pull out two-place verbs. Later, when the single primitive is used, these verbs will be defined already.

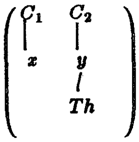
argument, y and z are identical arguments, and that the aspectual class is *activity*. The system returns the total TMI, $\theta_T(c)$.

$\theta_P(propel)$ =

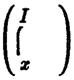


(7) *Apply the minimum TMI to the complete set C: Return to (6).* This applies the minimal thematic mapping over set C. Check results interactively with the user as critic. The minimal thematic mapping for a word in this set is:

$c \in C$ =



(8) *Apply Thematic Specialists to C:* These extract incorporated information from the entries. For example, in the definition of *throw*, we encounter after the head *propel*, the phrase *with motion of the arm.* Two Thematic specialists operate on this phrase, both for incorporation of the *Instrument* as well as a secondary *Theme*, or moving object, i.e. the arm. This knowledge is represented in the Thematic Mapping Index explicitly, as in



for the *Instrument.*

(9) *Update Primitive Set:* Add the words in C to the set P, forming a derived set, P_1.

(10) *Return to Step (1):* Repeat with a primitive selected from P_1.

At this point, the system can select a primitive or a derived primitive, such as *propel*, to do specialization over. Suppose we select *propel.* This will define the set containing *throw*, etc. the hierarchy being formed as a result will embed this tree within the tree formed from the previous run. We will discuss this process in more length in the final paper.

6. Related Research and Conclusion

In this paper we have tried to motivate a particularly rich lexical structure for dictionary entries. Given this representation in terms of the *Extended Aspect Calculus*, we presented a knowledge acquisition system that generates robust lexical structures from Machine Readable Dictionaries. The knowledge added to an entry includes a full specification of the argument types, selectional properties, the aspectual classification of the verb, and the thematically incorporated information. The information we have hand-coded is richer than that provided by the Longman, LDOCE ([Procter, 1978]), but it is quite feasible to automate the acquisition of their information with our system.[13]

We motivated the thematic mapping index as a useful way to generalize over argument structures and conceptual types. For example, this will be a convenient representation for lexical selection in the generation process (Cf. [Pustejovsky et al, 1987]).

There are several problems we have been unable to address. First, how does the system determine the direction of the search in the acquisition mode? We suggested some heuristics in section four, but this is still an open question. Another issue left open is how general the *thematic specialists* are and can be. Eventually, one would like such information-extractors to be generalizable over different MRDs.

Finally, there is the issue of how this work relates to the machine learning literature. The generalization performed in steps (5) and (7) to the entire word set constitute a conservative induction step, with an interactive credit assignment. The semantic model limits what a possible word type can be, and this in effect specializes the most general rule descriptions, increasing the number of maximally-specific specializations. A more automated lexicon construction algorithm is our current research goal. We will also compare with works such as [Haas and Hendrix, 1983], [Ballard and Stumberger, 1986], as well as [Anderson, 1986] and [Lebowitz, 1986]. As is, however, we hope the model here could act in conjunction with a system such as Amsler's [Amsler, 1980] for improved performance.

References

[AHD, 1983] AHD, *The American Heritage Dictionary*, Dell Publishing, New York, 1983.

[Amsler, 1980] Amsler, Robert, "The Structure of the Merriam Webster Pocket Dictionary". Ph.D. Thesis, University of Texas, Austin, Texas, 1980

[Atkin et al, 1986] Atkin, Beryl T, Judy Kegl, and Beth Levin, "Explicit and Implicit Information in Dictionaries, CSL Report 5, Princeton University, 1986.

[Bach, 1986] Bach, Emmon, "The Algebra of Events", in *Linguistics and Philosophy*, 1986.

[Chodorow et al,1985] Chodorow, Martin S, Roy J. Byrd, and George E. Heidorn, "Extracting Semantic Hierarchies from a Large On-Line dictionary", in Proceedings of the 23 Annual Meeting of the Association of Computational Linguistics", Chicago, Ill, 1985.n

[Cumming, 1986] Cumming, Sussana, "The Distribution of Lexical Information in Text Generation", presented for Workshop on Automating the Lexicon, Pisa, 1986.

[Dowty, 1979] Dowty, David R., Word Meaning and Montague Grammar. D. Reidel. Dordrecht, Holland, 1979.

[Ingria, 1986] Ingria, Robert, "Lexical Information for Parsing Systems: Points of Convergence and Divergence", Workshop on Automating the Lexicon, Pisa, 1986

[Jackendoff, 1972] Jackendoff, Ray, *Semanic Interpretation in Generative Grammar*, MIT Press, Cambridge. MA. 1972

[Pustejovsky, 1987] Pustejovsky, James, "The Extended Aspect Calculus", Submission to special issue of *Computational Linguistics*, 1987.

[Pustejovsky et al, 1987b] Pustejovsky, James and Sergei Nirenburg, "Lexical Selection in the Process of Generation, to appear in Proceedings of ACL, 1987b.