Clustering Lexical Patterns Obtained from a Text Corpus

Howard W. Beck and Balaji Kumar

IFAS Information Technologies
PO Box 110495
University of Florida
Gainesville, FL 32611
hwb@agen.ufl.edu

Abstract

A system for lexical acquisition is presented where word meanings are represented by clusters of phrase patterns obtained from analysis of a text corpus. A sample of cases, in the form of a concordance of phrases in which a particular word occurs in the text, is used for the basic analysis. Clustering techniques are used to group together cases having similar grammar and/or meaning. This view is that words obtain their meaning from the category describing this clustering of cases. This category is theory-based in that it contains a model to represent the word meaning at an abstract level, whereas the cases provide empirical evidence which confirm or disprove the model. A complex category evolves as more cases are encountered. Each new case matches to an existing category, or may dynamically alter existing categories as needed to account for the new case. An experimental system is presented which includes syntactic and semantic analysis of phrases obtained from text. It uses a hand-built lexicon and grammar to bootstrap a learning process. The ability to dynamically alter category structure through interpretation of new cases is shown as a way to build lexical structure semi-automatically.

Introduction

A lexicon must be designed on the basis of a theory of word meaning. The theory of word meaning being applied here is fashioned after work on category theory in developmental psychology. Basically, words are attached to categories. The problem is to describe the complex structure of categories. Initial work on category theory (Rosch and Mervis, 1975, Neisser 1987, Lakoff, 1987) shows that categories exhibit many properties, such as basic level, prototypes, family resemblance, default values, and cognitive models. Categories are dynamic, accommodating new evidence by altering their structure to interpret and explain the new evidence. Categories incorporate both theory, an abstract description of what it means to be a member of a category, and empirical data, inductions and generalizations over large collections of observations. These ideas are applied in the "theory-theory" of developmental psychology (Gopnik and Meltzof, 1997), which argues that children are basically scientists (the process of cognitive development in children is essentially the same as the processes by which scientists acquire knowledge through evolution of scientific theories). New observations are treated as empirical evidence which must be explained by existing theories. Theories may be overturned and replaced by new theories in the face of evidence which is incompatible with existing theories.

In the case of word meaning, the many uses and senses of a particular word are described by a category which contains classes (theories) describing (explaining) a set of cases in which a word is used. The category contains many overlapping subclasses to describe all these cases. The meaning of a word does not reduce to a single, simple definition (or a finite set of sense definitions), but is a dynamic representation based on a large number of cases and clusters over this set of cases. Such a representation can, through a process of case-based reasoning (Beck, 1991, Schank and Leake, 1989), handle new cases which may not map exactly to existing cases or definitions.

But such a complex knowledge base cannot be generated automatically from nothing. Before machine learning techniques can be used, an existing knowledge base must be constructed by hand (in gratitude to such manual methods, they are in most cases the only ones we have). Such a hand-constructed knowledge base is used to bootstrap the learning process. This has been done by several researchers , including (Poznanski and Sanfilippo, 1996) in which a machine readable dictionary (MRD) is refined by lexical acquisition from a text corpus. Of course, MRDs are constructed manually.

A clustering algorithm (Beck et al., 1994) uses case-based reasoning techniques to group cases of usage for a particular word into a category with sub-categories describing various word senses. The idea behind case-based reasoning is that we recognize a phrase containing a word by relating that phrase to previously encountered

cases of usage for that word. A theory of meaning for the particular word evolves over time by evolution of a category which can account for all the observed usage. The function of the clustering algorithm is to build this category.

This paper presents an example of the process of acquiring lexical data from analysis of a text corpus where the initial knowledge base is constructed by hand, but is extended semi-automatically by new cases extracted from the corpus.

Procedure

The text corpus used is a collection of 100 publications on home lawn care. In fact this project is part of an information retrieval system containing 4000 publications from the College of Agriculture at the University of Florida. This collection has not been processed except for tagging in SGML (which is not at a level fine enough for text analysis) and indexing through fulltext search engines.

The first step of analysis was hand construction of a basic lexicon, grammar, and concept taxonomy. The initial lexicon was generated by extracting a word list (a 4000 word vocabulary) from the fulltext search index which had been created for the lawn care publications. These were grouped into clusters having a common stem and identified initially with part-of-speech tags.

The grammar was initially generated by analysis of cases for words selected for study. Grammar is represented in a hierarchy of phrase patterns consisting of high-level, abstract patterns (of the S->NP,VP variety) using part-of-speech tags as categories, as well as low-level semantic patterns (S → <management_practice> <controls> <problem>) using labels from an initial concept taxonomy as categories. This taxonomy[1] was generated from a list of keywords which was created independently for use as an index to the home lawn care publications. The keywords were manually arranged into a generalization hierarchy.

For more detailed analysis of a particular word, a concordance of phrases containing the word was generated from the fulltext search index. These phrases were then parsed using an island chart parser that could generate a semantic representation of the phrase. For example, the phrases:

selective control of emerged grass weeds such as crabgrass (1a)

to control emerged summer annual grass weeds such as crabgrass (1b)

[1] The taxonomy can be seen at http://hammock.ifas.ufl.edu/thesaurus

are parsed syntactically as:

control[selective] (2a)
(weeds[emerged, grass, crabgrass])

control (2b)
(weeds[emerged, summer, annual,
grass, crabgrass])

where phrase heads are shown as predicates with arguments in parentheses and modifiers in brackets. The parser is capable of identifying local phrases in case a parse of the entire sentence is not possible.

Each phrase pattern used in parsing has an associated template which is filled from the elements of the pattern. For example, for the pattern:

<strategy> control of <weed>

there is a template:

control
strategy: 1
weed: 4

where the numbers indicate position of terms in the pattern. These numbers are substituted with actual values from a phrase which has been parsed using this pattern.

In such a way, semantic representations of 2a and 2b are created as:

control (3a)
strategy: selective
problem: weed
type: grass
stage: emerged
example: crabgrass

control (3b)
problem: weed
type: grass, annual
stage: emerged
season: summer
example: crabgrass

The semantic representations are incrementally clustered into the concept taxonomy through a matching process. The matching algorithm is based on a match between two graphs. The matching algorithm determines how the graphs are similar by matching them node-by-node and trying to find a class in the concept taxonomy which subsumes both nodes.

The semantic representations, such as 3a and 3b, are converted into graph form to simplify and generalize the

matching process. Conversion to a graph is achieved by creating a node for each term in the template. Links connect nodes for neighboring terms in the template.

Two graphs are systematically matched by first finding a node which appears in both graphs (if the two graphs are for phrases coming from the same concordance, such a node is easy to find since they share the word for which the concordance was generated). Using this node as a starting point, neighboring nodes in each graph are compared to see if they can be matched. The process continues in order to find a common sub-graph. Because the search proceeds from the initial common node along links to neighboring nodes, the search space is greatly reduced (as opposed to trying all combinations of the many possible node pairs that can be generated between two graphs).

For example, intersecting two graphs created from 3a and 3b produce the generalization:

control
  problem: weed
    type: grass
    stage: emerged
    example: crabgrass

Repeatedly applying this process across all cases in the concordance generates a clustering of phrase patterns for a given lexical entry. Each new phrase encountered causes this cluster to increase in size and complexity, improving its ability to interpret new phrases.

Results

Figure 1 shows an example clustering for the word "control". There were 62 cases of "control" occurring within the collection. The figure shows general groupings, but in fact there are even more associations than can be expressed in this simplified, two-dimensional

desired control

control period

biological control agent (2)
biological control agents (4)
additional control agents
biological control measures (2)
chemical control measures

disease control program
most nematode control treatments

IPM control tactics (4)
IPM control strategies
chemical control strategies
good cultural control strategy

biological control
biological controls (3)
biological pest control
the concept of biological control
chemical control
chemical controls
non-chemical controls
cultural controls (2)

pest control (2)
pest controls
effective pest control
oldest means of pest control
classic biological pest control
insect control
better insect control
mole cricket control
weed control
nutsedge control
turf nematode control
turfgrass diseases and their control

weed control guide

preventative controls for all pests
good controls for all nematode problems
best control of many insects and weeds
the control of Japanese beetle grubs.
biological control of turf pests
biocontrol of diseases
selective control of emerged grass weeds

to control emerged summer annual grass weeds
control insects
this option selectively controls fire ants
control another organism.
a selective herbicide controls certain plant species
nonselective herbicides control green plants regardless of species

pest to be controlled
chemical agents which control pests

Figure 1. A clustering of phrases for the word "control". Numbers in parentheses indicate that a phrase appeared more than once in the corpus.

figure (for example, the "<quality> control" patterns such as "effective control", "good control", "better control", "best control" also form a cluster).

When "control" occurs as the head of a noun phrase, it is most often used to describe either the mechanism ofcontrol (e.g. biological control) or the problem being controlled (e.g. pest control, or controls for all pests). As a verb, "control <problem>" patterns are most frequent with the problem being controlled appearing as the direct object.

Phrases such as "desired control" or "control period" which don't cluster closely with any of the other cases can be matched with clusters from other lexical entries (such as clusters for "desired" or "period").

General patterns can be abstracted from each of the major groupings shown in the figure. But within the groupings there are subtle variations among the cases. For example, "control agents", "control measures", "control programs", and "control treatments", introduce slight variations in meaning even though they can all be generalized to a "control <tool>" pattern.

An example of novel usage is illustrated by the phrase "non-selective herbicides control green plants regardless of species". Since all other cases of the verb "control" have <pest problem> as the direct object, then this usage forces "green plants" to be interpreted as a kind of pest problem, when in fact not all green plants are pests. The weight of evidence provided by the large number of cases in the existing clusters helps to support this interpretation.

Of course, the cluster shown here is reflective of the sub-domain of study (lawn care). Applying this technique to a wider corpus would encounter new uses of "control" (such as "remote control"). But it is possible that new uses could be related to one or more of the existing cases (such as "<mechanism> control"), dynamically extending of the existing clustering.

Related Work

Pustejovsky and Boguraev (1993), in a similar way, represent word meanings as dynamic processes. Rather than using static lists of word senses to enumerate the possible meanings of a word, they use a generative theory of lexical semantics which can extend coverage to novel uses. They use type coercion to account for situations where predicate arguments may have novel types (as in our "...control green plants..." example given above). Our system uses graph matching, induction, and reasoning from cases to accomplish similar results. One issue concerns reducing the size of the lexicon. In the generative lexicon, the number of patterns used to describe word meaning is greatly reduced. In our approach, the size of the lexicon can become quite large. We argue that this is a desirable feature, and that a large case-base and resulting cluster are required to cover the wide range of usage and provide a basis for reasoning to novel cases. We use an object-oriented database management systems to store large numbers of cases efficiently. Rather than slow down parsing, having many low-level cases can speed up processing by eliminating the need to derive an interpretation when an existing case may already do the job. Of course, when a word is well understood (has many cases) additional cases can become redundant, and then there is not a need to store every case encountered.

CRYSTAL (Soderland et al., 1995) uses learning-by-example techniques to induce a general concept-node definition from a set of minimally tagged phrase cases for a particular word. Although our approach is an attempt at unsupervised learning, it shares in common with CRYSTAL the need for some number of prior manually-generated data structures (both systems require a pre-existing concept taxonomy). But we argue that inducing a single (or a few) definition(s) which covers all and only the training set cases violates a fundamental characteristic of word meaning. The more cases of usage there are for a particular word, the greater the number of subtle variations in meaning and the less there is that all the cases have in common (hence, there would be little or nothing to induce as a generalization). These variations are what allow both the interpretation and generation of novel uses.

There have been few other attempts at applying case-based reasoning to lexical acquisition. Cardie (1993) shows that CBR can be used successfully to infer syntactic and semantic features of unknown words. It uses a k-nearest neighbor similarity metric based on numerical rankings using number of overlapping features to find similar cases. In our approach, no numerical methods are used, rather similarity between two concepts is defined by having common ancestors in a concept classification taxonomy.

Conclusions

It is shown that learning new lexical patterns can be accomplished in a semi-automatic fashion beginning with a hand-built knowledge base of lexical patterns, grammar rules, and categories which are extended through interpretation of new lexical patterns observed in a text corpus. An incremental conceptual clustering algorithm is used to achieve this dynamic learning. The program was demonstrated on an example domain.

The practical applications of this system are in information retrieval. Since it acts as repository of concepts occurring within the application domain, the resulting knowledge base provides a way of searching the document collection used in the text corpus with a high degree of precision and recall.

References

Beck, H. 1991. Language acquisition from cases. Proceedings of the DARPA Case-based Reasoning Workshop. Morgan Kaufmann, Inc. San Mateo, CA. pp. 159-169.

Beck, H. , T. Anwar, and S. Navathe. 1994. A conceptual clustering algorithm for database schema design. IEEE Transactions on Knowledge and Data Engineering. 6(3): 396-411.

Cardie, C.. 1993. A case-basd approach to knowledge acquisition for domain-specific sentence analysis. Proceedings of the Eleventh National Conference on Artificial Intelligence. AAAI Press/MIT Press. pp. 798-803.

Gopnik, A., and A. Meltzof. 1997. Words, thoughts, and theories. MIT Press. Cambridge, MA.

Lakoff, G.. 1987. Women, fire, and dangerous things. University of Chicago Press. Chicago, IL.

Lenat, D., and R. Guha. 1990. Building large knowledge-based systems. Addison-Wesley, Reading, MA.

Neisser, U. ed. 1987. Concepts and conceptual development: Ecological and intellectual factors in categorization, Cambridge University, Cambridge, MA.

Poznanski, V. and A. Sanfilippo. 1996. Detecting dependencies between semantic verb subclasses and subcategorization frames in text corpora. In Corpus processing for lexical acquistion. B. Boguraev and J.Pustejovsky (eds.). MIT Press. Cambridge, MA.

Pustejovsky, J. and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. Artificial Intelligence. 63:193-223.

Rosch, E. and C. B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology. 7:573-605.

Schank, R.C. and D. B. Leake. 1989. Creativity and learning in a case-based explainer. Artificial Intelligence. 40:353-385.

Soderland, S., D. Fisher, J. Aseltine, and W. Lehnert. 1995. CRYSTAL: Inducing a conceptual dictionary. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp. 1314-1319.