Using Terminological Knowledge Representation Languages to Manage Linguistic Resources

Pamela W. Jordan
Intelligent Systems Program
University of Pittsburgh
Pittsburgh PA 15260
jordan@isp.pitt.edu

Abstract

I examine how terminological languages can be used to manage linguistic data during NL research and development. In particular, I consider the lexical semantics task of characterizing semantic verb classes and show how the language can be extended to flag inconsistencies in verb class definitions, identify the need for new verb classes, and identify appropriate linguistic hypotheses for a new verb's behavior.

1 Introduction

Problems with consistency and completeness can arise when writing a wide-coverage grammar or analyzing lexical data since both tasks involve working with large amounts of data. Since terminological knowledge representation languages have been valuable for managing data in other applications such as a software information system that manages a large knowledge base of plans (Devanbu and Litman, 1991), it is worthwhile considering how these languages can be used in linguistic data management tasks. In addition to inheritance, terminological systems provide a criterial semantics for links and automatic classification which inserts a new concept into a taxonomy so that it directly links to concepts more general than it and more specific than it (Woods and Schmolze, 1992).

Terminological languages have been used in NLP applications for lexical representation (Burkert, 1995), and grammar representation (Brachman and Schmolze, 1991), and to assist in the acquisition and maintenance of domain specific lexical semantics knowledge (Ayuso et al., 1987). Here I explore additional linguistic data management tasks. In particular I examine how a terminological language such as Classic (Brachman et al., 1991) can assist a lexical semanticist with the management of verb classes. In conclusion, I discuss ways in which terminological languages can be used during grammar writing.

Consider the tasks that confront a lexical semanticist. The regular participation of verbs belonging to a particular semantic class in a limited number of syntactic alternations is crucial in lexical semantics. A popular research direction assumes that the syntactic behavior of a verb is systematically influenced by its meaning (Levin, 1993; Hale and Keyser, 1987) and that any set of verbs whose members pattern together with respect to syntactic alternations should form a semantically coherent class (Levin, 1993). Once such a class is identified, the meaning component that the member verbs share can be identified. This gives further insight into lexical representation for the words in the class (Levin, 1993).

Terminological languages can support three important functions in this domain. First, the process of representing the system in a taxonomic logic can serve as a check on the rigor and precision of the original account. Once the account is represented, the terminological system can flag inconsistencies. Second, the classifier can identify an existing verb class that might explain an unassigned verb's behavior. That is, given a set of syntactically analyzed sentences that exemplify the syntactic alternations allowed and disallowed for that verb, the classifier will provide appropriate linguistic hypotheses. Third, the classifier can identify the need for new verb classes by flagging verbs that are not members of any existing, defined verb classes. Together, these functions provide tools for the lexical semanticist that are potentially very useful.

The second and third of these three functions can be provided in two steps: (1) classifying each alternation for a particular verb according to the type of semantic mapping allowed for the verb and its arguments; and (2) either identifying the verb class that has the given pattern of classified alternations or using the pattern to form the definition of a new verb class.

2 Sentence Classification

The usual practice in investigating the alternation patterning of a verb is to construct example sentences in which simple, illustrative noun phrases are used as arguments of a verb. The sentences in (1)

exemplify two familiar alternations of *give*.

(1) a. John gave Mary a book

b. John gave a book to Mary.

Such sentences exemplify an alternation that belongs to the alternation pattern of their verb.[1] I will call this the *alternation type* of the test sentence.

To determine the alternation type of a test sentence, the sentence must be syntactically analyzed so that its grammatical functions (e.g. subject, object) are marked. Then, given semantic feature information about the words filling those grammatical functions (GFs), and information about the possible argument structures for the verb in the sentence and the semantic feature restrictions on these arguments, it is possible to find the argument structures appropriate to the input sentence. Consider the sentences and descriptions shown below for *pour*:

(2) a. [Mary$_{subj}$] poured [Tina$_{obj}$] [a glass of milk$_{io}$].

b. [Mary$_{subj}$] poured [a glass of milk$_{obj}$] for [Tina$_{ppo}$].

pour$_1$: subj → agent$_{[volitional]}$
obj → recipient$_{[volitional]}$
io → patient$_{[liquid]}$

pour$_2$: subj → agent$_{[volitional]}$
obj → patient$_{[liquid]}$
ppo → recipient$_{[volitional]}$

Given the semantic type restrictions and the GFs, *pour*$_1$ describes (2a) and *pour*$_2$, (2b). The mapping from the GFs to the appropriate argument structure is similar to lexical rules in the LFG syntactic theory except that here I semantically type the arguments. To indicate the alternation types for these sentences, I call sentence (2a) a benefactive-ditransitive and sentence (2b) a benefactive-transitive.

Classifying a sentence by its alternation type requires linguistic and world knowledge. World knowledge is used in the definitions of nouns and verbs in the lexicon and describes high-level entities, such as events, and animate and inanimate objects. Properties (such as LIQUID) are used to define specialized entities. For example, the property NON-CONSUMABLE (SMALL CAPITALS indicate Classic concepts in my implementation) specializes a LIQUID-ENTITY to define PAINT and distinguish it from WATER, which has the property that it is CONSUMABLE. Specialized EVENT entities are used in the definition of verbs in the lexicon and represent the argument structures for the verbs.

The linguistic knowledge needed to support sentence classification includes the definitions of (1) verb types such as intransitive, transitive and ditransitive; (2) verb definitions; and (3) concepts that define the links between the GFs and verb argument structures as represented by events.

[1]In the examples that I will consider, and in most examples used by linguists to test alternation patterns, there will only be one verb; this is the verb to be tested.

Verb types (SUBCATEGORIZATIONS) are defined according to the GFs found in the sentence. For example, (2a) classifies as DITRANSITIVE and (2b) as a specialized TRANSITIVE with a PP. Once the verb type is identified, verb definitions (VERBs) are needed to provide the argument structures. A VERB can have multiple senses which are instances of EVENTs, for example the verb "pour" can have the senses *pour* or *prepare*, with the required arguments shown below.[2] Note that *pour*$_1$ and *pour*$_2$ in (2) are subcategorizations of *prepare*.

pour: pourer$_{[volitional]}$
pouree$_{[inanimate-container]}$
poured$_{[inanimate-substance]}$

prepare: preparer$_{[volitional]}$
prepareе$_{[liquid]}$
prepared$_{[volitional]}$

For a sentence to classify as a particular ALTERNATION, a legal linking must exist between an EVENT and the SUBCATEGORIZATION. Linking involves restricting the fillers of the GFs in the SUBCATEGORIZATION to be the same as the arguments in an EVENT. In Classic, the **same-as** restriction is limited so that either both attributes must be filled already with the same instance or the concept must already be known as a LEGAL-LINKING. Because of this I created a test (written in LISP) to identify a LEGAL-LINKING. The test inputs are the sentence predicate and GF fillers arranged in the order of the event arguments against which they are to be tested. A linking is legal when at least one of the events associated with the verb can be linked in the indicated way, and all the required arguments are filled.

Once a sentence passes the linking test, and classifies as a particular ALTERNATION, a rule associated with the ALTERNATION classifies it as a specialization of the concept. This causes the EVENT arguments to be filled with the appropriate GF fillers from the SUBCATEGORIZATION. A side-effect of the alternation classification is that the EVENT classifies as a specialized EVENT and indicates which sense of the verb is used in the sentence.

3 Semantic Class Classification

The semantic class of the verb can be identified once the example sentences are classified by their alternation type. Specialized VERB-CLASSes are defined by their good and bad alternations. Note that VERB defines one verb whereas VERB-CLASS describes a set of verbs (e.g. spray/load class). Which ALTERNATIONs are associated with a VERB-CLASS is a matter of linguistic evidence; the linguist discovers these associations by testing examples for grammaticality. To assist in this task, I provide two tests, **have-instances-of** and **have-no-instances-of**.

[2]For generality in the implementation, I use arg$_1$ ... arg$_n$ for all event definitions instead of agent ... patient or preparer ... preparee.

The **have-instances-of** test for an ALTERNATION searches a corpus of good sentences or bad sentences and tests whether at least one instance of the specified ALTERNATION, for example a benefactive-ditransitive, is present.

A bad sentence with all the required verb arguments will classify as an ALTERNATION despite the ungrammatical syntactic realization, while a bad sentence with missing required arguments will only classify as a SUBCATEGORIZATION. The **have-no-instances-of** test for a SUBCATEGORIZATION searches a corpus of bad sentences and tests whether at least one instance of the specified SUBCATEGORIZATION, for example TRANSITIVE, is present as the most specific classification.

4 Discussion

The ultimate test of this approach is in how well it will scale up. The linguist may choose to add knowledge as it is needed or may prefer to do this work in batches. To support the batch approach, it may be useful to extract detailed subcategorization information from English learner's dictionaries. Also it will be necessary to decide what semantic features are needed to restrict the fillers of the argument structures. Finally, there is the problem of collecting complete sets of example sentences for a verb. In general, a corpus of tagged sentences is inadequate since it rarely includes negative examples and is not guaranteed to exhibit the full range of alternations. In applications where a domain specific corpus is available (e.g. the Kant MT project (Mitamura et al., 1993)), the full range of relevant alternations is more likely. However, the lack of negative examples still poses a problem and would require the project linguist to create appropriate negative examples or manually adjust the class definitions for further differentiation.

While I have focused on a lexical research tool, an area I will explore in future work is how classification could be used in grammar writing. One task for which a terminological language is appropriate is flagging inconsistent rules. When writing and maintaining a large grammar, inconsistent rules is one type of grammar writing bug that occurs. For example, the following three rules are inconsistent since feature_1 of NP and feature_1 of VP would not unify in rule 1 given the values assigned in 2 and 3.

1) S → NP VP
   <NP feature_1 > = <VP feature_1 >
2) NP → det N
   <N feature_1 > = +
   <NP> = <N>
3) VP → V
   <V feature_1 > = −
   <VP> = <V>

5 Conclusion

I have shown how a terminological language, such as Classic, can be used to manage lexical semantics data during analysis with two minor extensions. First, a test to identify LEGAL-LINKINGs is necessary since this cannot be directly expressed in the language and second, set membership tests, **have-instances-of** and **have-no-instances-of** are necessary since this type of expressiveness is not provided in Classic. While the solution of several knowledge acquisition issues would result in a friendlier tool for a linguistics researcher, the tool still performs a useful function.

References

Damaris M. Ayuso, Varda Shaked, and Ralph Weischedel. 1987. An environment for acquiring semantic information. In *Proceedings of 25th ACL*, pages 32–40.

Ronald J. Brachman and James Schmolze. 1991. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216.

Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lori A. Resnik. 1991. Living with CLASSIC: When and how to use a KL-ONE-like language. In John F. Sowa, editor, *Principles of Semantic Networks*, pages 401–456. Morgan Kaufmann, San Mateo, CA.

Gerrit Burkert. 1995. Lexical semantics and terminological knowledge representation. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*. Cambridge University Press.

Premkumar Devanbu and Diane J. Litman. 1991. Plan-based terminological reasoning. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *KR'91: Principles of Knowledge Representation and Reasoning*, pages 128–138. Morgan Kaufmann, San Mateo, CA.

K. L. Hale and S. J. Keyser. 1987. A view from the middle. Center for Cognitive Science, MIT. Lexicon Project Working Papers 10.

B. Levin. 1993. *English verb classes and alternations: a preliminary investigation.* University of Chicago Press.

T. Mitamura, E. Nyberg, and J. Carbonell. 1993. Automated corpus analysis and the acquisition of large, multi-lingual knowledge bases for MT. In *Proceedings of TMI-93.*

William A. Woods and James G. Schmolze. 1992. The KL-ONE family. In Fritz Lehmann, editor, *Semantic Networks in Artificial Intelligence*, pages 133–177. Pergamon Press, Oxford.