Seeding lexical semantics: resources using parsed corpora

Alastair Butler, Stephen Wright Horn and Iku Nagasaki
National Institute for Japanese Language and Linguistics (NINJAL), Tokyo

This paper describes new techniques that take advantage of the Keyaki Treebank (a parsed corpus of modern Japanese), together with a program that indexes semantic dependency relationships between syntactic constituents onto nodes of parsed trees. The trees and the indexing are new resources that find application in at least four specific areas:

1) the graphic display of semantic dependencies,
2) a metric for evaluating semantic dependencies,
3) the ability to search for semantic dependencies, and
4) the seeding of frames for annotation in lexical semantics.

The Keyaki Treebank describes three basic kinds of dependency between grammatical elements: modification, argumenthood, and antecedent inheritance. The ways in which these relationships are defined actually determines the annotation used in the Keyaki Treebank.

Modification is typically a local relationship, but can also obtain across conjuncts, and sometimes long-distance relationships must be described. Attachment height takes care of local modification and modification across conjuncts, and a simple indexing accounts for long-distance modification.

The second type of dependency is that between argument and head. For overt arguments, this is typically local, but upstairs argument antecedence for downstairs subject gaps is frequently attested, and this dependency is calculated by a mechanism called control. The control mechanism employs an accessibility hierarchy and structural relations to define the relationships that obtain without the use of indexing. Across the Board (ATB) extraction is a relationship that obtains across clausal conjuncts, and includes argument-head relations. This calculation too is carried out without resorting to indexing.

The last kind of dependency is antecedence inheritance, and its calculation for non-local modifiers and arguments has already been mentioned above. However, both overt and null pronouns have their reference resolved in the Keyaki Treebank through the addition of “binding information” (a simple and robust method of establishing a link between an element and the domain).

Accordingly, in the Keyaki Treebank the computation of relationships of antecedent inheritance is carried out by four mechanisms:

1) by an argument acting as a local binder or by ATB,
2) by an argument filled through control,
3) by an argument filled by long-distance indexing, or
4) by an overt or null pronoun that finds an accessible antecedent through binding information.

The combination of these four mechanisms, together with the contribution of functional and lexical

elements, allows the generation of a predicate logic description of the semantic composition of sentence meaning. A full display might be an expression of a predicate logic language. Another method is a tree where local dependencies between overtly annotated elements are read configurationally, and non-local relations and relations calculated through control and ATB are indexed onto nodes.

There are still other displays that are easier to grasp and visually process. Furthermore, there are more abstract levels of meaning that can be added to the corpus to capture lexical meanings and their interdependencies. Finally, the definition of abstract dependencies makes them amenable to a statistical analysis for the first time. This paper discusses these possibilities, and the results of some practical experiments towards their realisation.

One of the graphic tools used in FrameNet displays semantic dependencies in a very intuitive way, that can be used, for example, by native speakers to check the accuracy of parsed corpora, or by learners of a target language to grasp the compositional elements of sentence interpretation. A translation of the Keyaki Treebank indexing to the character-indexed format of the FrameNet data model allows the extraction from Keyaki trees of the relations to be displayed in the FrameNet interface. This application is called a “Dependency array.”

As an extension of the practice of generating indexes for non-local dependencies onto nodes in a tree, a program can be developed to express the dependencies involved in a given predication in a specific context as a set of attributes of the predicate itself. This technique can be used in the seeding of a lexical semantic resource that assigns semantic roles (frame elements) and a specific semantic frame to each attestation of a predicate. This process enables systematic sense disambiguation to be carried out as well. The actual annotation process takes the form of adding these assignments to a pre-generated description in the node label of the predicate in the tree. Lexical entries in a dictionary can thereby be linked through their senses to specific attestations, and the expression of a role in a frame can be traced to realisations with different predicates.

The calculation of non-local and abstract dependencies (expressed in trees as secondary edges) in a robust and principled way allows for them to be expressed in a variety of forms. One of these forms is compatible with Tiger Search, a tool which allows searches for secondary edges of various types, paired with any variety of structural conditions. Development of the technique would allow surveys of the distribution of null pronouns, for example, to be conducted on a full set of corpus data.

Finally, the development of a metric to evaluate parsed trees on the basis of the completeness and coherence of the semantic dependencies encoded therein is a contribution both to quality control in corpus development, and to various practical applications such as the evaluation of machine translations, the measurement of similarity in utterance/response pairs in dialogue systems, and the generation of paraphrase texts. Encoding semantic dependencies in such a way as to measure alignments between parsings is a new method that makes these applications possible.