Sometime between the 7th and 4th centuries BCE, the Indian grammarian Pāṇini wrote a famous treatise on Sanskrit grammar, the Āṣṭādhyāyī (‘8 books’), a treatise that has been called “one of the greatest monuments of human intelligence” (Bloomfield, 1933, 11). The work describes the linguistics of the Sanskrit language in the form of 3959 sutras, each very efficiently (since it had to be memorized!) expressing part of a formal rule system that brilliantly prefigured modern mechanisms of formal language theory (Penn and Kiparsky, 2012). One set of rules, relevant to our discussion in this chapter, describes the kārakas, semantic relationships between a verb and noun arguments, roles like agent, instrument, or destination. Pāṇini’s work was the earliest we know of that tried to understand the linguistic realization of events and their participants. This task of understanding participants and their relationship to events—being able to answer the question “Who did what to whom” (and perhaps also “when and where”)—is a central question of natural language understanding.

Let’s move forward 2.5 millenia to the present and consider the very mundane goal of understanding text about a purchase of stock by XYZ Corporation. This purchasing event could take on a wide variety of surface forms. In the following sentences we see that it could be described by a verb (sold, bought) or a noun (purchase), and that XYZ Corp can be the syntactic subject (of bought), the indirect object (of sold), or in a genitive or noun compound relation (with the noun purchase) despite having notationally the same role in all of them:

• XYZ corporation bought the stock.
• They sold the stock to XYZ corporation.
• The stock was bought by XYZ corporation.
• The purchase of the stock by XYZ corporation...
• The stock purchase by XYZ corporation...

In this chapter we introduce a level of representation that lets us capture the commonality between these sentences. We will be able to represent the fact that there was a purchase event, that the participants in this event were XYZ Corp and some stock, and that XYZ Corp played a specific role, the role of acquiring the stock.

We call this shallow semantic representation level semantic roles. Semantic roles are representations that express the abstract role that arguments of a predicate can take in the event; these can be very specific, like the BUYER, abstract like the AGENT, or super-abstract (the PROTO-AGENT). These roles can both represent general semantic properties of the arguments and also express their likely relationship to the syntactic role of the argument in the sentence. AGENTS tend to be the subject of

---

1 Figure shows a birch bark manuscript from Kashmir of the Rupavatra, a grammatical textbook based on the Sanskrit grammar of Panini. Image from the Wellcome Collection.
an active sentence, THEMES the direct object, and so on. These relations are codified in databases like PropBank and FrameNet. We’ll introduce **semantic role labeling**, the task of assigning roles to the constituents or phrases in sentences. We’ll also discuss **selectional restrictions**, the semantic sortal restrictions or preferences that each individual predicate can express about its potential arguments, such as the fact that the theme of the verb *eat* is generally something edible. Along the way, we’ll describe the various ways these representations can help in language understanding tasks like question answering and machine translation.

### 18.1 Semantic Roles

Consider how in Chapter 14 we represented the meaning of arguments for sentences like these:

(18.1) Sasha broke the window.
(18.2) Pat opened the door.

A neo-Davidsonian event representation of these two sentences would be

\[
\exists e, x, y \Breaking(e) \land \Breaker(e, Sasha) \\
\land \BrokenThing(e, y) \land \Window(y) \\
\exists e, x, y \Opening(e) \land \Opener(e, Pat) \\
\land \OpenedThing(e, y) \land \Door(y)
\]

In this representation, the roles of the subjects of the verbs *break* and *open* are *Breaker* and *Opener* respectively. These **deep roles** are specific to each event; *Breaking* events have *Breakers*, *Opening* events have *Openers*, and so on.

If we are going to be able to answer questions, perform inferences, or do any further kinds of natural language understanding of these events, we’ll need to know a little more about the semantics of these arguments. *Breakers* and *Openers* have something in common. They are both volitional actors, often animate, and they have direct causal responsibility for their events.

**Thematic roles** are a way to capture this semantic commonality between *Breakers* and *Eaters*. We say that the subjects of both these verbs are *agents*. Thus, *agent* is the thematic role that represents an abstract idea such as volitional causation. Similarly, the direct objects of both these verbs, the *BrokenThing* and *OpenedThing*, are both prototypically inanimate objects that are affected in some way by the action.

The semantic role for these participants is **theme**.

Although thematic roles are one of the oldest linguistic models, as we saw above, their modern formulation is due to Fillmore (1968) and Gruber (1965). Although there is no universally agreed-upon set of roles, Figs. 18.1 and 18.2 list some thematic roles that have been used in various computational papers, together with rough definitions and examples. Most thematic role sets have about a dozen roles, but we’ll see sets with smaller numbers of roles with even more abstract meanings, and sets with very large numbers of roles that are specific to situations. We’ll use the general term **semantic roles** for all sets of roles, whether small or large.
18.2 Diathesis Alternations

The main reason computational systems use semantic roles is to act as a shallow meaning representation that can let us make simple inferences that aren’t possible from the pure surface string of words, or even from the parse tree. To extend the earlier examples, if a document says that Company A acquired Company B, we’d like to know that this answers the query Was Company B acquired? despite the fact that the two sentences have very different surface syntax. Similarly, this shallow semantics might act as a useful intermediate language in machine translation.

Semantic roles thus help generalize over different surface realizations of predicate arguments. For example, while the AGENT is often realized as the subject of the sentence, in other cases the THEME can be the subject. Consider these possible realizations of the thematic arguments of the verb break:

(18.3) \textit{John} broke the window.
\begin{align*}
\text{AGENT} & \quad \text{THEME} \\
\text{John} & \quad \text{broke} \quad \text{the} \quad \text{window}.
\end{align*}

(18.4) \textit{John} broke the window with a rock.
\begin{align*}
\text{AGENT} & \quad \text{THEME} \quad \text{INSTRUMENT} \\
\text{John} & \quad \text{broke} \quad \text{the} \quad \text{window} \quad \text{with} \quad \text{a} \quad \text{rock}.
\end{align*}

(18.5) The rock broke the window.
\begin{align*}
\text{INSTRUMENT} & \quad \text{THEME} \\
\text{The} \quad \text{rock} & \quad \text{broke} \quad \text{the} \quad \text{window}.
\end{align*}

(18.6) The window broke.
\begin{align*}
\text{THEME} \\
\text{The} \quad \text{window} & \quad \text{broke}.
\end{align*}

(18.7) The window was broken by John.
\begin{align*}
\text{THEME} \quad \text{AGENT} \\
\text{The} \quad \text{window} & \quad \text{was} \quad \text{broken} \quad \text{by} \quad \text{John}.
\end{align*}
These examples suggest that *break* has (at least) the possible arguments *AGENT*, *THEME*, and *INSTRUMENT*. The set of thematic role arguments taken by a verb is often called the *thematic grid*, θ-grid, or *case frame*. We can see that there are (among others) the following possibilities for the realization of these arguments of *break*:

AGENT/Subject, THEME/Object
AGENT/Subject, THEME/Object, INSTRUMENT/PP with
INSTRUMENT/Subject, THEME/Object
THEME/Subject

It turns out that many verbs allow their thematic roles to be realized in various syntactic positions. For example, verbs like *give* can realize the *THEME* and *GOAL* arguments in two different ways:

(18.8) a. Doris gave the book to Cary.
AGENT THEME GOAL
b. Doris gave Cary the book.
AGENT GOAL THEME

These multiple argument structure realizations (the fact that *break* can take *AGENT*, *INSTRUMENT*, or *THEME* as subject, and *give* can realize its *THEME* and *GOAL* in either order) are called *verb alternations* or *diathesis alternations*. The alternation we showed above for *give*, the *dative alternation*, seems to occur with particular semantic classes of verbs, including “verbs of future having” (advance, allocate, offer, owe), “send verbs” (forward, hand, mail), “verbs of throwing” (kick, pass, throw), and so on. Levin (1993) lists for 3100 English verbs the semantic classes to which they belong (47 high-level classes, divided into 193 more specific classes) and the various alternations in which they participate. These lists of verb classes have been incorporated into the online resource VerbNet (Kipper et al., 2000), which links each verb to both WordNet and FrameNet entries.

### 18.3 Semantic Roles: Problems with Thematic Roles

Representing meaning at the thematic role level seems like it should be useful in dealing with complications like diathesis alternations. Yet it has proved quite difficult to come up with a standard set of roles, and equally difficult to produce a formal definition of roles like *AGENT*, *THEME*, or *INSTRUMENT*.

For example, researchers attempting to define role sets often find they need to fragment a role like *AGENT* or *THEME* into many specific roles. Levin and Rappaport Hovav (2005) summarize a number of such cases, such as the fact there seem to be at least two kinds of *INSTRUMENTS*, *intermediary* instruments that can appear as subjects and *enabling* instruments that cannot:

(18.9) a. The cook opened the jar with the new gadget.
b. The new gadget opened the jar.

(18.10) a. Shelly ate the sliced banana with a fork.
b. *The fork ate the sliced banana.

In addition to the fragmentation problem, there are cases in which we’d like to reason about and generalize across semantic roles, but the finite discrete lists of roles don’t let us do this.
Finally, it has proved difficult to formally define the thematic roles. Consider the AGENT role; most cases of AGENTS are animate, volitional, sentient, causal, but any individual noun phrase might not exhibit all of these properties.

These problems have led to alternative semantic role models that use either many fewer or many more roles.

The first of these options is to define generalized semantic roles that abstract over the specific thematic roles. For example, PROTO-AGENT and PROTO-PATIENT are generalized roles that express roughly agent-like and roughly patient-like meanings. These roles are defined, not by necessary and sufficient conditions, but rather by a set of heuristic features that accompany more agent-like or more patient-like meanings. Thus, the more an argument displays agent-like properties (being volitionally involved in the event, causing an event or a change of state in another participant, being sentient or intentionally involved, moving) the greater the likelihood that the argument can be labeled a PROTO-AGENT. The more patient-like the properties (undergoing change of state, causally affected by another participant, stationary relative to other participants, etc.), the greater the likelihood that the argument can be labeled a PROTO-PATIENT.

The second direction is instead to define semantic roles that are specific to a particular verb or a particular group of semantically related verbs or nouns.

In the next two sections we describe two commonly used lexical resources that make use of these alternative versions of semantic roles. PropBank uses both proto-roles and verb-specific semantic roles. FrameNet uses semantic roles that are specific to a general semantic idea called a frame.

18.4 The Proposition Bank

The Proposition Bank, generally referred to as PropBank, is a resource of sentences annotated with semantic roles. The English PropBank labels all the sentences in the Penn TreeBank; the Chinese PropBank labels sentences in the Penn Chinese TreeBank. Because of the difficulty of defining a universal set of thematic roles, the semantic roles in PropBank are defined with respect to an individual verb sense. Each sense of each verb thus has a specific set of roles, which are given only numbers rather than names: Arg0, Arg1, Arg2, and so on. In general, Arg0 represents the PROTO-AGENT, and Arg1, the PROTO-PATIENT. The semantics of the other roles are less consistent, often being defined specifically for each verb. Nonetheless there are some generalization; the Arg2 is often the benefactive, instrument, attribute, or end state, the Arg3 the start point, benefactive, instrument, or attribute, and the Arg4 the end point.

Here are some slightly simplified PropBank entries for one sense each of the verbs agree and fall. Such PropBank entries are called frame files; note that the definitions in the frame file for each role (“Other entity agreeing”, “Extent, amount fallen”) are informal glosses intended to be read by humans, rather than being formal definitions.

(18.11) agree.01
Arg0: Agreer  
Arg1: Proposition  
Arg2: Other entity agreeing

Ex1: [Arg0 The group] agreed [Arg1 it wouldn’t make an offer].  
Ex2: [ArgM-TMP Usually] [Arg0 John] agrees [Arg2 with Mary] [Arg1 on everything].

(18.12) **fall.01**
Arg1: Logical subject, patient, thing falling  
Arg2: Extent, amount fallen  
Arg3: start point  
Arg4: end point, end state of arg1

Ex1: [Arg1 Sales] fell [Arg4 to $25 million] [Arg3 from $27 million].  
Ex2: [Arg1 The average junk bond] fell [Arg2 by 4.2%].

Note that there is no Arg0 role for *fall*, because the normal subject of *fall* is a PROTO-PATIENT.

The PropBank semantic roles can be useful in recovering shallow semantic information about verbal arguments. Consider the verb *increase*:

(18.13) **increase.01** “go up incrementally”
Arg0: causer of increase  
Arg1: thing increasing  
Arg2: amount increased by, EXT, or MNR  
Arg3: start point  
Arg4: end point

A PropBank semantic role labeling would allow us to infer the commonality in the event structures of the following three examples, that is, that in each case Big Fruit Co. is the AGENT and the price of bananas is the THEME, despite the differing surface forms.

(18.14) [Arg0 Big Fruit Co. ] increased [Arg1 the price of bananas].  
(18.15) [Arg1 The price of bananas] was increased again [Arg0 by Big Fruit Co. ]  
(18.16) [Arg1 The price of bananas] increased [Arg2 5%].

PropBank also has a number of non-numbered arguments called ArgMs. (ArgM-TMP, ArgM-LOC, etc) which represent modification or adjunct meanings. These are relatively stable across predicates, so aren’t listed with each frame file. Data labeled with these modifiers can be helpful in training systems to detect temporal, location, or directional modification across predicates. Some of the ArgM’s include:

- **TMP** when? yesterday evening, now  
- **LOC** where? at the museum, in San Francisco  
- **DIR** where to/from? down, to Bangkok  
- **MNR** how? clearly, with much enthusiasm  
- **PRP/CAU** why? because ..., in response to the ruling  
- **REC** themselves, each other  
- **ADV** miscellaneous  
- **PRD** secondary predication ...ate the meat raw

**NomBank** While PropBank focuses on verbs, a related project, NomBank (Meyers et al., 2004) adds annotations to noun predicates. For example the noun agreement in *Apple’s agreement with IBM* would be labeled with Apple as the Arg0 and IBM as
the Arg2. This allows semantic role labelers to assign labels to arguments of both verbal and nominal predicates.

18.5 FrameNet

While making inferences about the semantic commonalities across different sentences with *increase* is useful, it would be even more useful if we could make such inferences in many more situations, across different verbs, and also between verbs and nouns. For example, we’d like to extract the similarity among these three sentences:

(18.17) [Arg1 The price of bananas] increased [Arg2 5%].
(18.18) [Arg1 The price of bananas] rose [Arg2 5%].
(18.19) There has been a [Arg2 5%] rise [Arg1 in the price of bananas].

Note that the second example uses the different verb *rise*, and the third example uses the noun rather than the verb *rise*. We’d like a system to recognize that the *price of bananas* is what went up, and that 5% is the amount it went up, no matter whether the 5% appears as the object of the verb *increased* or as a nominal modifier of the noun *rise*.

The FrameNet project is another semantic-role-labeling project that attempts to address just these kinds of problems (Baker et al. 1998, Fillmore et al. 2003, Fillmore and Baker 2009, Ruppenhofer et al. 2016). Whereas roles in the PropBank project are specific to an individual verb, roles in the FrameNet project are specific to a frame.

What is a frame? Consider the following set of words:

*reservation, flight, travel, buy, price, cost, fare, rates, meal, plane*

There are many individual lexical relations of hyponymy, synonymy, and so on between many of the words in this list. The resulting set of relations does not, however, add up to a complete account of how these words are related. They are clearly all defined with respect to a coherent chunk of common-sense background information concerning air travel.

We call the holistic background knowledge that unites these words a frame (Fillmore, 1985). The idea that groups of words are defined with respect to some background information is widespread in artificial intelligence and cognitive science, where besides frame we see related works like a model (Johnson-Laird, 1983), or even script (Schank and Abelson, 1977).

A frame in FrameNet is a background knowledge structure that defines a set of frame-specific semantic roles, called frame elements, and includes a set of predicates that use these roles. Each word evokes a frame and profiles some aspect of the frame and its elements. The FrameNet dataset includes a set of frames and frame elements, the lexical units associated with each frame, and a set of labeled example sentences. For example, the change_position_on_a_scale frame is defined as follows:

This frame consists of words that indicate the change of an Item’s position on a scale (the Attribute) from a starting point (Initial_value) to an end point (Final_value).

Some of the semantic roles (frame elements) in the frame are defined as in Fig. 18.3. Note that these are separated into core roles, which are frame specific, and
non-core roles, which are more like the Arg-M arguments in PropBank, expressed more general properties of time, location, and so on.

Core Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTRIBUTE</td>
<td>The ATTRIBUTE is a scalar property that the ITEM possesses.</td>
</tr>
<tr>
<td>DIFFERENCE</td>
<td>The distance by which an ITEM changes its position on the scale.</td>
</tr>
<tr>
<td>FINAL_STATE</td>
<td>A description that presents the ITEM’s state after the change in the ATTRIBUTE’s value as an independent predication.</td>
</tr>
<tr>
<td>FINAL_VALUE</td>
<td>The position on the scale where the ITEM ends up.</td>
</tr>
<tr>
<td>INITIAL_STATE</td>
<td>A description that presents the ITEM’s state before the change in the ATTRIBUTE’s value as an independent predication.</td>
</tr>
<tr>
<td>INITIAL_VALUE</td>
<td>The initial position on the scale from which the ITEM moves away.</td>
</tr>
<tr>
<td>Item</td>
<td>The entity that has a position on the scale.</td>
</tr>
<tr>
<td>Value_range</td>
<td>A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate.</td>
</tr>
</tbody>
</table>

Some Non-Core Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>The length of time over which the change takes place.</td>
</tr>
<tr>
<td>Speed</td>
<td>The rate of change of the Value.</td>
</tr>
<tr>
<td>Group</td>
<td>The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way.</td>
</tr>
</tbody>
</table>

Figure 18.3 The frame elements in the change_position_on_a_scale frame from the FrameNet Labelers Guide (Ruppenhofer et al., 2016).

Here are some example sentences:

(18.20) [ITEM Oil] rose [ATTRIBUTE in price] [DIFFERENCE by 2%].

(18.21) [ITEM It] has increased [FINAL_STATE to having them 1 day a month].

(18.22) [ITEM Microsoft shares] fell [FINAL_VALUE to 7 5/8].

(18.23) [ITEM Colon cancer incidence] fell [DIFFERENCE by 50%] [GROUP among men].

(18.24) a steady increase [INITIAL_VALUE from 9.5] [FINAL_VALUE to 14.3] [ITEM in dividends]

(18.25) a [DIFFERENCE 5%] [ITEM dividend] increase...

Note from these example sentences that the frame includes target words like rise, fall, and increase. In fact, the complete frame consists of the following words:

VERBS: dwindle, move, soar, escalation, shift
climb, explode, mushroom, swell, explosion, tumble
advance, fall, reach, triple, fluctuation
decline, increase, skyrocket, decrease, rise

ADVERBS: increasingly, gain, growth, increasingly, gain, growth

NOUNS: hike, decline, increase, shift, slide, drop

FrameNet also codes relationships between frames, allowing frames to inherit from each other, or representing relations between frames like causation (and generalizations among frame elements in different frames can be representing by inheritance as well). Thus, there is a Cause_change_of_position_on_a_scale frame that is linked to the Change_of_position_on_a_scale frame by the cause relation, but that adds an AGENT role and is used for causative examples such as the following:
Together, these two frames would allow an understanding system to extract the common event semantics of all the verbal and nominal causative and non-causative usages.

FrameNets have also been developed for many other languages including Spanish, German, Japanese, Portuguese, Italian, and Chinese.

18.6 Semantic Role Labeling

Semantic role labeling (sometimes shortened as SRL) is the task of automatically finding the semantic roles of each argument of each predicate in a sentence. Current approaches to semantic role labeling are based on supervised machine learning, often using the FrameNet and PropBank resources to specify what counts as a predicate, define the set of roles used in the task, and provide training and test sets.

Recall that the difference between these two models of semantic roles is that FrameNet (18.27) employs many frame-specific frame elements as roles, while PropBank (18.28) uses a smaller number of numbered argument labels that can be interpreted as verb-specific labels, along with the more general ARGMM labels. Some examples:

(18.27) [You] can’t [blame] [the program] [for being unable to identify it]

(18.28) [The San Francisco Examiner] issued [a special edition] [yesterday]

18.6.1 A Feature-based Algorithm for Semantic Role Labeling

A simplified feature-based semantic role labeling algorithm is sketched in Fig. 18.4. Feature-based algorithms—from the very earliest systems like (Simmons, 1973)—begin by parsing, using broad-coverage parsers to assign a parse to the input string. Figure 18.5 shows a parse of (18.28) above. The parse is then traversed to find all words that are predicates.

For each of these predicates, the algorithm examines each node in the parse tree and uses supervised classification to decide the semantic role (if any) it plays for this predicate. Given a labeled training set such as PropBank or FrameNet, a feature vector is extracted for each node, using feature templates described in the next subsection. A 1-of-N classifier is then trained to predict a semantic role for each constituent given these features, where N is the number of potential semantic roles plus an extra NONE role for non-role constituents. Any standard classification algorithms can be used. Finally, for each test sentence to be labeled, the classifier is run on each relevant constituent.

Instead of training a single-stage classifier as in Fig. 18.5, the node-level classification task can be broken down into multiple steps:

1. **Pruning:** Since only a small number of the constituents in a sentence are arguments of any given predicate, many systems use simple heuristics to prune unlikely constituents.

2. **Identification:** a binary classification of each node as an argument to be labeled or a NONE.
function SEMANTICROLELABEL(words) returns labeled tree

parse ← PARSE(words)

for each predicate in parse do
    for each node in parse do
        featurevector ← EXTRACTFEATURES(node, predicate, parse)
        CLASSIFYNODE(node, featurevector, parse)

Figure 18.4 A generic semantic-role-labeling algorithm. CLASSIFYNODE is a 1-of-N classifier that assigns a semantic role (or NONE for non-role constituents), trained on labeled data such as FrameNet or PropBank.

Figure 18.5 Parse tree for a PropBank sentence, showing the PropBank argument labels. The dotted line shows the path feature NP↑S↑VP↓VBD for ARG0, the NP-SBJ constituent The San Francisco Examiner.

3. **Classification**: a 1-of-N classification of all the constituents that were labeled as arguments by the previous stage

The separation of identification and classification may lead to better use of features (different features may be useful for the two tasks) or to computational efficiency.

**Global Optimization**

The classification algorithm of Fig. 18.5 classifies each argument separately (‘locally’), making the simplifying assumption that each argument of a predicate can be labeled independently. This assumption is false; there are interactions between arguments that require a more ‘global’ assignment of labels to constituents. For example, constituents in FrameNet and PropBank are required to be non-overlapping. More significantly, the semantic roles of constituents are not independent. For example PropBank does not allow multiple identical arguments; two constituents of the same verb cannot both be labeled ARG0.

Role labeling systems thus often add a fourth step to deal with global consistency across the labels in a sentence. For example, the local classifiers can return a list of possible labels associated with probabilities for each constituent, and a second-pass Viterbi decoding or re-ranking approach can be used to choose the best consensus label. Integer linear programming (ILP) is another common way to choose a solution that conforms best to multiple constraints.
Features for Semantic Role Labeling

Most systems use some generalization of the core set of features introduced by Gildea and Jurafsky (2000). Common basic features templates (demonstrated on the NP-SBJ constituent The San Francisco Examiner in Fig. 18.5) include:

- The governing predicate, in this case the verb issued. The predicate is a crucial feature since labels are defined only with respect to a particular predicate.
- The phrase type of the constituent, in this case, NP (or NP-SBJ). Some semantic roles tend to appear as NPs, others as S or PP, and so on.
- The headword of the constituent, Examiner. The headword of a constituent can be computed with standard head rules, such as those given in Chapter 10 in Fig. ???. Certain headwords (e.g., pronouns) place strong constraints on the possible semantic roles they are likely to fill.
- The headword part of speech of the constituent, NNP.
- The path in the parse tree from the constituent to the predicate. This path is marked by the dotted line in Fig. 18.5. Following Gildea and Jurafsky (2000), we can use a simple linear representation of the path, NP↑S↓VP↓VBD. ↑ and ↓ represent upward and downward movement in the tree, respectively. The path is very useful as a compact representation of many kinds of grammatical function relationships between the constituent and the predicate.
- The voice of the clause in which the constituent appears, in this case, active (as contrasted with passive). Passive sentences tend to have strongly different linkings of semantic roles to surface form than do active ones.
- The binary linear position of the constituent with respect to the predicate, either before or after.
- The subcategorization of the predicate, the set of expected arguments that appear in the verb phrase. We can extract this information by using the phrase-structure rule that expands the immediate parent of the predicate; VP → VBD NP PP for the predicate in Fig. 18.5.
- The named entity type of the constituent.
- The first words and the last word of the constituent.

The following feature vector thus represents the first NP in our example (recall that most observations will have the value NONE rather than, for example, ARG0, since most constituents in the parse tree will not bear a semantic role):

ARG0: [issued, NP, Examiner, NNP, NP↑S↓VP↓VBD, active, before, VP → NP PP, ORG, The, Examiner]

Other features are often used in addition, such as sets of n-grams inside the constituent, or more complex versions of the path features (the upward or downward halves, or whether particular nodes occur in the path).

It’s also possible to use dependency parses instead of constituency parses as the basis of features, for example using dependency parse paths instead of constituency paths.

18.6.2 A Neural Algorithm for Semantic Role Labeling

The standard neural algorithm for semantic role labeling is based on the bi-LSTM IOB tagger introduced in Chapter 9, which we’ve seen applied to part-of-speech tagging and named entity tagging, among other tasks. Recall that with IOB tagging,