Chapter 22
Information Extraction

I am the very model of a modern Major-General,
I've information vegetable, animal, and mineral,
I know the kings of England, and I quote the fights historical
From Marathon to Waterloo, in order categorical...
Gilbert and Sullivan, Pirates of Penzance

Imagine that you are an analyst with an investment firm that tracks airline stocks. You're given the task of determining the relationship (if any) between airline announcements of fare increases and the behavior of their stocks on the following day. Historical data about stock prices is easy to come by, but what about the information about airline announcements? To do a reasonable job on this task, you would need to know at least the name of the airline, the nature of the proposed fare hike, the dates of the announcement and possibly the response of other airlines. Fortunately, this information resides in archives of news articles reporting on airline's actions, as in the following recent example.

> Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Of course, distilling information like names, dates and amounts from naturally occurring text is a non-trivial task. This chapter presents a series of techniques that can be used to extract limited kinds of semantic content from text. This process of **information extraction** (IE) turns the unstructured information embedded in texts into structured data. More concretely, information extraction is an effective way to populate the contents of a relational database. Once the information is encoded formally, we can apply all the capabilities provided by database systems, statistical analysis packages and other forms of decision support systems to address the problems we're trying to solve.

Information extraction

As we proceed through this chapter, we'll see that robust solutions to IE problems are actually clever combinations of techniques we've seen earlier in the book. In particular, the finite-state methods described in Chs. 2 and 3, the probabilistic models introduced in Chs. 4 through 6 and the syntactic chunking methods from Ch. 13 form the core of most current approaches to information extraction. Before diving into the details of how these techniques are applied, let's quickly introduce the major problems in IE and how they can be approached.

The first step in most IE tasks is to detect and classify all the proper names mentioned in a text — a task generally referred to as **named entity recognition** (NER). Not

Named entity recognition

surprisingly, what constitutes a proper name and the particular scheme used to classify them is application-specific. Generic NER systems tend to focus on finding the names of people, places and organizations that are mentioned in ordinary news texts; practical applications have also been built to detect everything from the names of genes and proteins (Settles, 2005) to the names of college courses (McCallum, 2005).

Named entity mentions

Our introductory example contains 13 instances of proper names, which we'll refer to as **named entity mentions**, which can be classified as either organizations, people, places, times or amounts.

Having located all of the mentions of named entities in a text, it is useful to link, or cluster, these mentions into sets that correspond to the entities behind the mentions. This is the task of **reference resolution**, which we introduced in Ch. 21, and is also an important component in IE. In our sample text, we would like to know that the *United Airlines* mention in the first sentence and the *United* mention in the third sentence refer to the same real world entity. This general reference resolution problem also includes anaphora resolution as a sub-problem. In this case, determining that the two uses of *it* refer to *United Airlines* and *United* respectively.

Relation detection and classification

The task of **relation detection and classification** is to find and classify semantic relations among the entities discovered in a given text. In most practical settings, the focus of relation detection is on small fixed sets of binary relations. Generic relations that appear in standard system evaluations include family, employment, part-whole, membership, and geospatial relations. The relation detection and classification task is the one that most closely corresponds to the problem of populating a relational database. Relation detection among entities is also closely related to the problem of discovering semantic relations among words introduced in Ch. 20.

Our sample text contains 3 explicit mentions of generic relations: *United* is a part of *UAL*, *American Airlines* is a part of *AMR* and *Tim Wagner* is an employee of *American Airlines*. Domain-specific relations from the airline industry would include the fact that *United* serves *Chicago*, *Dallas*, *Denver* and *San Francisco*.

In addition to knowing about the entities in a text and their relation to one another, we might like to find and classify the events in which the entities are participating; this is the problem of **event detection and classification**. In our sample text, the key events are the fare increase by *United* and the ensuing increase by *American*. In addition, there are several events reporting these main events as indicated by the two uses of *said* and the use of *cite*. As with entity recognition, event detection brings with it the problem of reference resolution; we need to figure out which of the many event mentions in a text refer to the same event. In our running example, the events referred to as *the move* and *the increase* in the second and third sentences are the same as the *increase* in the first sentence.

Event detection and classification

Temporal expression recognition
Temporal analysis

The problem of figuring out when the events in a text happened and how they relate to each other in time raises the twin problems of **temporal expression recognition** and **temporal analysis**. Temporal expression detection tells us that our sample text contains the temporal expressions *Friday* and *Thursday*. Temporal expressions include date expressions such as days of the week, months, holidays, etc., as well as relative expressions including phrases like *two days from now* or *next year*. They also include expressions for clock times such as *noon* or *3:30PM*.

The overall problem of **temporal analysis** is to map temporal expressions onto

specific calendar dates or times of day and then to use those times to situate events in time. It includes the following subtasks.

- Fixing the temporal expressions with respect to an anchoring date or time, typically the dateline of the story in the case of news stories;
- Associating temporal expressions with the events in the text;
- Arranging the events into a complete and coherent timeline.

In our sample text, the temporal expressions *Friday* and *Thursday* should be anchored with respect to the dateline associated with the article itself. We also know that *Friday* refers to the time of United's announcement, and *Thursday* refers to the time that the fare increase went into effect (i.e. the Thursday immediately preceding the Friday). Finally, we can use this information to produce a timeline where United's announcement follows the fare increase and American's announcement follows both of those events. Temporal analysis of this kind is useful in nearly any NLP application that deals with meaning, including question answering, summarization and dialogue systems.

Template-filling

Finally, many texts describe stereotypical situations that recur with some frequency in the domain of interest. The task of **template-filling** is to find documents that evoke such situations and then fill the slots in templates with appropriate material. These slot-fillers may consist of text segments extracted directly from the text, or they may consist of concepts that have been inferred from text elements via some additional processing (times, amounts, entities from an ontology, etc.).

Our airline text is an example of this kind of stereotypical situation since airlines are often attempting to raise fares and then waiting to see if competitors follow along. In this situation, we can identify *United* as a lead airline that initially raised its fares, $6 as the amount by which fares are being raised, *Thursday* as the effective date for the fare increase, and *American* as an airline that followed along. A filled template from our original airline story might look like the following.

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

The following sections will review current approaches to each of these problems in the context of generic news text. Section 22.5 then describes how many of these problems arise in the context of procecessing biology texts.

22.1 Named Entity Recognition

Named entity

The starting point for most information extraction applications is the detection and classification of the named entities in a text. By **named entity**, we simply mean anything that can be referred to with a proper name. This process of **named entity recognition**

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains and automobiles

Figure 22.1 A list of generic named entity types with the kinds of entities they refer to.

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

Figure 22.2 Named entity types with examples.

refers to the combined task of finding spans of text that constitute proper names and then classifying the entities being referred to according to their type.

Generic news-oriented NER systems focus on the detection of things like people, places, and organizations. Figures 22.1 and 22.2 provide lists of typical named entity types with examples of each. Specialized applications may be concerned with many other types of entities, including commercial products, weapons, works of art, or as we'll see in Sec. 22.5, proteins, genes and other biological entities. What these applications all share is a concern with proper names, the characteristic ways that such names are signaled in a given language or genre, and a fixed set of categories of entities from a domain of interest.

By the way that names are signaled, we simply mean that names are denoted in a way that sets them apart from ordinary text. For example, if we're dealing with standard English text, then two adjacent capitalized words in the middle of a text are likely to constitute a name. Further, if they are preceded by a *Dr.* or followed by an *MD*, then it is likely that we're dealing with a person. In contrast, if they are preceded by *arrived in* or followed by *NY* then we're probably dealing with a location. Note that these signals include facts about the proper names as well as their surrounding contexts.

The notion of a named entity is commonly extended to include things that aren't entities per se, but nevertheless have practical importance and do have characteristic signatures that signal their presence; examples include dates, times, named events and other kinds of **temporal expressions**, as well as measurements, counts, prices and other kinds of **numerical expressions**. We'll consider some of these later in Sec. 22.3.

Temporal expressions
Numerical expressions

Let's revisit the sample text introduced earlier with the named entities marked (with TIME and MONEY used to to mark the temporal and monetary expressions).

> Citing high fuel prices, [$_{ORG}$ United Airlines] said [$_{TIME}$ Friday] it has increased fares by [$_{MONEY}$ $6] per round trip on flights to some cities also served by lower-cost carriers. [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Figure 22.3 Common categorical ambiguities associated with various proper names.

[$_{PERS}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{ORG}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{LOC}$ Washington] for what may well be his last state visit.
In June, [$_{GPE}$ Washington] passed a primary seatbelt law.
The [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

Figure 22.4 Examples of type ambiguities in the use of the name *Washington*.

matched the move, spokesman [$_{PERS}$ Tim Wagner] said. [$_{ORG}$ United], a unit of [$_{ORG}$ UAL Corp.], said the increase took effect [$_{TIME}$ Thursday] and applies to most routes where it competes against discount carriers, such as [$_{LOC}$ Chicago] to [$_{LOC}$ Dallas] and [$_{LOC}$ Denver] to [$_{LOC}$ San Francisco].

As shown, this text contains 13 mentions of named entities including 5 organizations, 4 locations, 2 times, 1 person, and 1 mention of money. The 5 organizational mentions correspond to 4 unique organizations, since *United* and *United Airlines* are distinct mentions that refer to the same entity.

22.1.1 Ambiguity in Named Entity Recognition

Named entity recognition systems face two types of ambiguity. The first arises from the fact the same name can refer to different entities of the same type. For example, *JFK* can refer to the former president or his son. This is basically a reference resolution problem and approaches to resolving this kind of ambiguity are discussed in Ch. 21.

The second source of ambiguity arises from the fact that identical named entity mentions can refer to entities of completely different types. For example, in addition to people, *JFK* might refer to the airport in New York, or to any number of schools, bridges and streets around the United States. Some examples of this kind of cross-type confusion are given in Figures 22.3 and 22.4.

Notice that some of the ambiguities shown in Fig. 22.3 are completely coincidental. There is no relationship between the financial and organizational uses of the name *IRA* — they simply arose coincidentally as acronyms from different sources (*Individual Retirement Account* and *International Reading Association*). On the other hand, the organizational uses of *Washington* and *Downing St.* are examples of a LOCATION-FOR-ORGANIZATION **metonymy**, as discussed in Ch. 19.

22.1.2 NER as Sequence Labeling

The standard way to approach the problem of named entity recognition is as a word-by-word sequence labeling task, where the assigned tags capture both the boundary and the type of any detected named entities. Viewed in this light, named entity recognition

looks very much like the problem of syntactic base-phrase chunking. In fact, the dominant approach to NER is based on the same statistical sequence labeling techniques introduced in Ch. 5 for part of speech tagging and Ch. 13 for syntactic chunking.

In the sequence labeling approach to NER, classifiers are trained to label the tokens in a text with tags that indicate the presence of particular kinds of named entities. This approach makes use of the same style of IOB encoding employed for syntactic chunking. Recall that in this scheme an I is used to label tokens *inside* of a chunk, B is used to mark the beginning of a chunk, and O labels tokens outside any chunk of interest. Consider the following sentence from our running example.

(22.1) [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PERS}$ Tim Wagner] said.

This bracketing notation provides us with the extent and the type of the named entities in this text. Fig. 22.5 shows a standard word-by-word IOB-style tagging that captures the same information. As with syntactic chunking, the tagset for such an encoding consists of 2 tags for each entity type being recognized, plus 1 for the O tag outside any entity, or $(2 \times N) + 1$ tags.

Having encoded our training data with IOB tags, the next step is to select a set of features to associate with each input example (i.e. each of the tokens to be labeled in Fig. 22.5). These features should be plausible predictors of the class label and should be easily and reliably extractable from the source text. Recall that such features can be based not only on characteristics of the token to be classified, but also on the text in a surrounding window as well.

Words	Label
American	B$_{ORG}$
Airlines	I$_{ORG}$
,	O
a	O
unit	O
of	O
AMR	B$_{ORG}$
Corp.	I$_{ORG}$
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B$_{PERS}$
Wagner	I$_{PERS}$
said	O
.	O

Figure 22.5 IOB encoding.

Fig. 22.6 gives a list of standard features employed in state-of-the-art named entity recognition systems. We've seen many of these features before in the context of part-of-speech tagging and syntactic base-phrase chunking. Several, however, are particularly important in the context of NER. The **shape feature** feature includes the usual upper case, lower case and capitalized forms, as well as more elaborate patterns designed to capture expressions that make use of numbers (*A9*), punctuation (*Yahoo!*) and atypical case alternations (*eBay*). It turns out that this feature by itself accounts for a considerable part of the success of NER systems for English news text. And as we'll see in Sec. 22.5, shape features are also particularly important in recognizing names of proteins and genes in biological texts. Fig. 22.7 describes some commonly employed shape feature values.

Shape feature

The **presence in a named entity list** feature can be very predictive. Extensive lists of names for all manner of things are available from both publicly available and commercial sources. Lists of place names, called **gazetteers**, contain millions of entries for all manner of locations along with detailed geographical, geologic and political in-

Gazetteers

Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or N-grams occurring in the surrounding context.

Figure 22.6 Features commonly used in training named entity recognition systems.

Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

Figure 22.7 Selected shape features.

formation.[1] The United States Census Bureau provides extensive lists of first names and surnames derived from its decadal census in the U.S.[2] Similar lists of corporations, commercial products, and all manner of things biological and mineral are also available from a variety of sources.

This feature is typically implemented as a binary vector with a bit for each available kind of name list. Unfortunately, such lists can be difficult to create and maintain, and their usefulness varies considerably based on the named entity class. It appears that gazetteers can be quite effective, while extensive lists of persons and organizations are not nearly as beneficial (Mikheev et al., 1999).

Finally, features based on the presence of **predictive words and N-grams** in the context window can also be very informative. When they are present, preceding and following titles, honorifics, and other markers such as *Rev.*, *MD* and *Inc.* can accurately indicate the class of an entity. Unlike name lists and gazetteers, these lists are relatively short and stable over time and are therefore easy to develop and maintain.

The relative usefulness of any of these features, or combination of features, depends to a great extent on the application, genre, media, language and text encoding. For example, shape features, which are critical for English newswire texts, are of little use with materials transcribed from spoken text via automatic speech recognition, materials gleaned from informally edited sources such as blogs and discussion forums, and for character-based languages like Chinese where case information isn't available. The set of features given in Fig. 22.6 should therefore be thought of as only a starting point for

[1] www.geonames.org

[2] www.census.gov

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Figure 22.8 Simple word-by-word feature encoding for NER.

any given application.

Once an adequate set of features has been developed, they are extracted from a representative training set and encoded in a form appropriate to train a machine learning-based sequence classifier. A standard way of encoding these features is to simply augment our earlier IOB scheme with more columns. Fig. 22.8 illustrates the result of adding part-of-speech tags, syntactic base-phrase chunk tags, and shape information to our earlier example.

Given such a training set, a sequential classifier can be trained to label new sentences. As with part-of-speech tagging and syntactic chunking, this problem can be cast either as Markov-style optimization using HMMs or MEMMs as described in Ch. 6, or as a multi-way classification task deployed as a sliding-window labeler as described in Ch. 13. Figure Fig. 22.9 illustrates the operation of such a sequence labeler at the point where the token *Corp.* is next to be labeled. If we assume a context window that includes the 2 preceding and following words, then the features available to the classifier are those shown in the boxed area. Fig. 22.10 summarizes the overall sequence labeling approach to creating a NER system.

22.1.3 Evaluating Named Entity Recognition

The familiar metrics of **recall**, **precision** and F_1 **measure** introduced in Ch. 13 are used to evaluate NER systems. Recall that recall is the ratio of the number of correctly labeled responses to the total that should have been labeled; precision is the ratio of the number of correctly labeled responses to the total labeled. The F-measure (van Rijsbergen, 1975) provides a way to combine these two measures into a single metric.

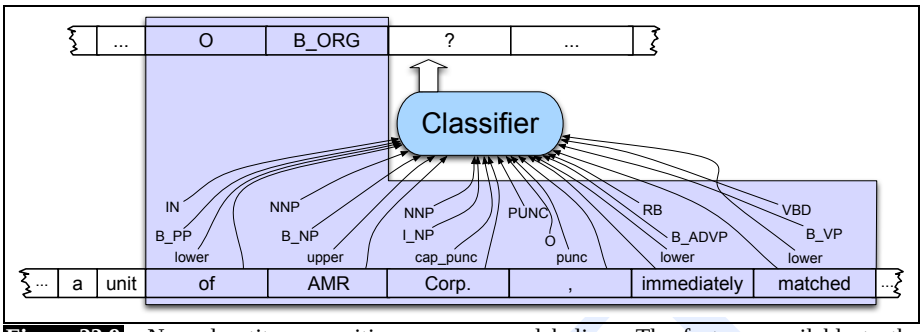

Figure 22.9 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

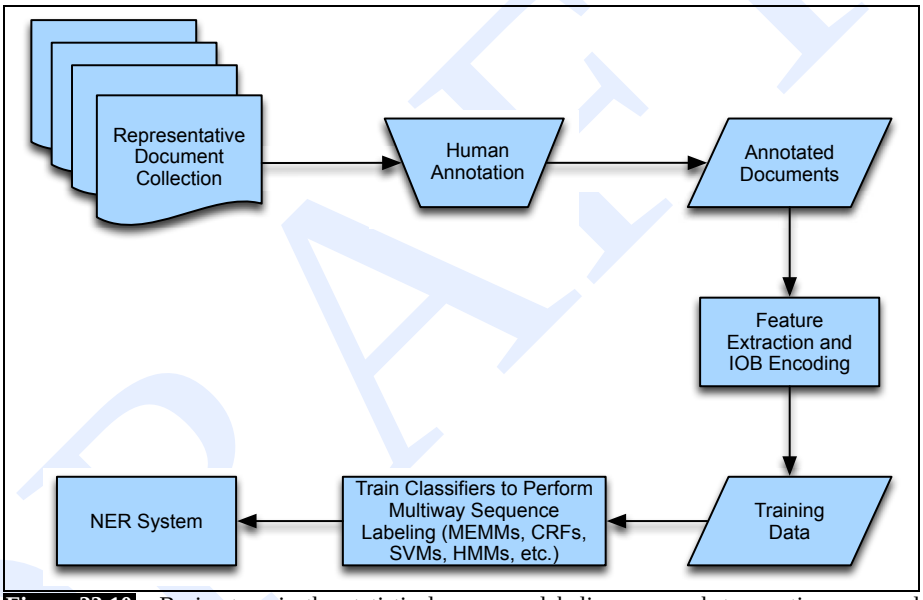

Figure 22.10 Basic steps in the statistical sequence labeling approach to creating a named entity recognition system.

The F-measure is defined as:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{22.2}$$

The β parameter is used to differentially weight the importance of recall and precision, based perhaps on the needs of an application. Values of $\beta > 1$ favor recall, while values of $\beta < 1$ favor precision. When $\beta = 1$, precision and recall are equally balanced; this is sometimes called $F_{\beta=1}$ or just F_1:

$$F_1 = \frac{2PR}{P+R} \tag{22.3}$$

As with syntactic chunking, it is important to distinguish the metrics used to measure performance at the application level from those used during training. At the appli-

cation level, recall and precision are measured with respect to the actual named entities detected. On the other hand, with an IOB encoding scheme the learning algorithms are attempting to optimize performance at the tag level. Performance at these two levels can be quite different; since the vast majority of tags in any given text are outside any entity, simply emitting an O tag for every token gives fairly high tag-level performance.

High-performing systems at recent standardized evaluations have entity level F-measures around .92 for PERSONS and LOCATIONS, and around .84 for ORGANIZATIONS (Sang and De Meulder, 2003).

22.1.4 Practical NER Architectures

Commercial approaches to NER are often based on pragmatic combinations of lists, rules and supervised machine learning(Jackson and Moulinier, 2002). One common approach is to make repeated passes over a text allowing the results of one pass to influence the next. The stages typically first involve the use of rules that have extremely high precision but low recall. Subsequent stages employ more error-prone statistical methods that take the output of the first pass into account.

1. First use high-precision rules to tag unambiguous entity mentions;
2. Then search for sub-string matches of the previously detected names using probabilistic string matching metrics (as described in Ch. 19).
3. Consult application-specific name lists to identify likely name entity mentions from the given domain.
4. Finally, apply probabilistic sequence labeling techniques that make use of the tags from previous stages as additional features.

The intuition behind this staged approach is two-fold. First, some of the entity mentions in a text will be more clearly indicative of a given entity's class than others. Second, once an unambiguous entity mention is introduced into a text, it is likely that subsequent shortened versions will refer to the same entity (and thus the same type of entity).

22.2 Relation Detection and Classification

Next on our list of tasks is the ability to discern the relationships that exist among the entities detected in a text. To see what this means, let's return to our sample airline text with all the entities marked.

> Citing high fuel prices, [$_{ORG}$ United Airlines] said [$_{TIME}$ Friday] it has increased fares by [$_{MONEY}$ $6] per round trip on flights to some cities also served by lower-cost carriers. [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PERS}$ Tim Wagner] said. [$_{ORG}$ United], a unit of [$_{ORG}$ UAL Corp.], said the increase took effect [$_{TIME}$ Thursday] and applies to most routes where it competes against discount carriers, such as [$_{LOC}$ Chicago] to [$_{LOC}$ Dallas] and [$_{LOC}$ Denver] to [$_{LOC}$ San Francisco].

This text stipulates a set of relations among the named entities mentioned within it. We know, for example, that *Tim Wagner* is a spokesman for *American Airlines*,

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Figure 22.11 Semantic relations with examples and the named entity types they involve.

that *United* is a unit of *UAL Corp.*, and that *American* is a unit of *AMR*. These are all binary relations that can be seen as instances of more generic relations such as **part-of** or **employs** that occur with fairly high frequency in news-style texts. Fig. 22.11 shows a list of generic relations of the kind used in recent standardized evaluations.[3] More domain-specific relations that might be extracted include the notion of an airline route. For example, from this text we can conclude that United has routes to Chicago, Dallas, Denver and San Francisco.

These relations correspond nicely to the model-theoretic notions we introduced in Ch. 17 to ground the meanings of the logical forms. That is, a relation consists of set of ordered tuples over elements of a domain. In most standard information extraction applications, the domain elements correspond either to the named entities that occur in the text, to the underlying entities that result from co-reference resolution, or to entities selected from a domain ontology. Fig. 22.12 shows a model-based view of the set of entities and relations that can be extracted from our running example. Notice how this model-theoretic view subsumes the NER task as well; named entity recognition corresponds to the identification of a class of unary relations.

22.2.1 Supervised Learning Approaches to Relation Analysis

Supervised machine learning approaches to relation detection and classification follow a scheme that should be familiar by now. Texts are annotated with relations chosen from a small fixed set by human analysts. These annotated texts are then used to train systems to reproduce similar annotations on unseen texts. Such annotations indicate the text spans of the two arguments, the roles played by each argument and the type of the relation involved.

The most straightforward approach breaks the problem down into two sub-tasks: detecting when a relation is present between two entities and then classifying any detected relations. In the first stage, a classifier is trained to make a binary decision as to whether or not a given pair of named entities participate in a relation. Positive examples are extracted directly from the annotated corpus, while negative examples are generated from within-sentence entity pairs that are not annotated with a relation.

[3] http://www.nist.gov/speech/tests/ace/

Domain	$\mathscr{D} = \{a,b,c,d,e,f,g,h,i\}$
United, UAL, American Airlines, AMR	a,b,c,d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f,g,h,i
Classes	
United, UAL, American and AMR are organizations	$Org = \{a,b,c,d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver and San Francisco are places	$Loc = \{f,g,h,i\}$
Relations	
United is a unit of UAL American is a unit of AMR	$PartOf = \{\langle a,b\rangle, \langle c,d\rangle\}$
Tim Wagner works for American Airlines	$OrgAff = \{\langle c,e\rangle\}$
United serves Chicago, Dallas, Denver and San Francisco	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Figure 22.12 A model-based view of the relations and entities in our sample text.

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

Figure 22.13 Finding and classifying the relations among entities in a text.

In the second phase, a classifier is trained to label the relations that exist between candidate entity pairs. As discussed in Ch. 6, techniques such as decision trees, naive Bayes or MaxEnt handle multiclass labeling directly. Binary approaches based on discovering separating hyperplanes such as SVMs solve multiclass problems by employing a one-versus-all training paradigm. In this approach, a sets of classifiers are trained where each classifier is trained on one label as the positive class and all the other labels as the negative class. Final classification is performed by passing each instance to be labeled to all of the classifiers and then choosing the label from the classifier with the most confidence, or returning a rank ordering over the positively responding classifiers. Fig. 22.13 illustrates the basic approach for finding and classifying relations among the named entities within a discourse unit.

As with named entity recognition, the most important step in this process is to identify surface features that will be useful for relation classification (Zhou et al., 2005). The first source of information to consider are **features of the named entities** themselves.

- Named entity types of the two candidate arguments
- Concatenation of the two entity types
- Head words of the arguments
- Bag of words from each of the arguments

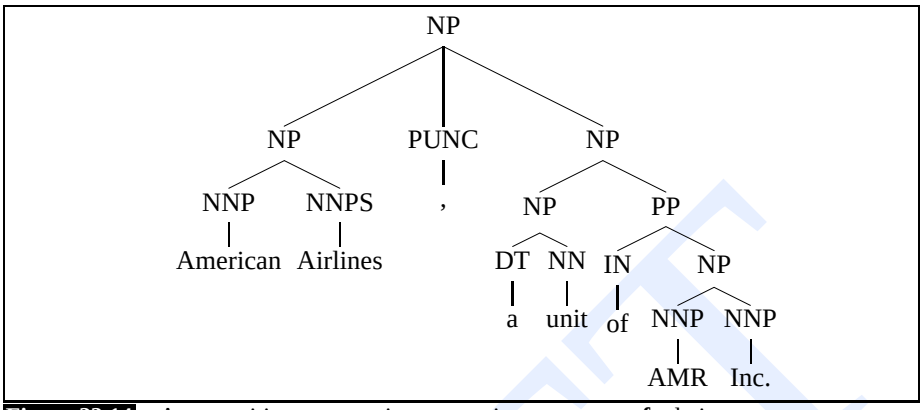

Figure 22.14 An appositive construction expressing an **a-part-of** relation.

The next set of features are derived from **the words in the text** being examined. It is useful to think of these features as being extracted from three locations: the text between the two candidate arguments, a fixed window before the first argument, and a fixed window after the second argument. Given these locations, the following word-based features have proven to be useful.

- The bag of words and bag of bigrams between the entities
- Stemmed versions of the same
- Words and stems immediately preceding and following the entities
- Distance in words between the arguments
- Number of entities between the arguments

Finally, **the syntactic structure** of a sentence can signal many of the relationships among any entities contained within it. The following features can be derived from various levels of syntactic analysis including base-phrase chunking, dependency parsing and full constituent parsing.

- Presence of particular constructions in a constituent structure
- Chunk base-phrase paths
- Bags of chunk heads
- Dependency-tree paths
- Constituent-tree paths
- Tree distance between the arguments

One method of exploiting parse trees is to create detectors that signal the presence of particular syntactic constructions and then associate binary features with those detectors. As an example of this, consider the sub-tree shown in Fig. 22.14 that dominates the named entities *American* and *AMR Inc.* The *NP* construction that dominates these two entities is called an appositive construction and is often associated with both **part-of** and **a-kind-of** relations in English. A binary feature indicating the presence of this construction can be useful in detecting these relations.

This method of feature extraction relies on a certain amount of a priori linguistic analysis to identify those syntactic constructions that may be useful predictors of cer-

Entity-based features	
Entity_1 type	ORG
Entity_1 head	*airlines*
Entity_2 type	PERS
Entity_2 head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity_1	NONE
Word(s) after Entity_2	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Figure 22.15 Sample of features extracted while classifying the <American Airlines, Tim Wagner> tuple.

tain classes. An alternative method is to automatically encode certain aspects of tree structures as feature values and allow the machine learning algorithms to determine which values are informative for which classes. One simple and effective way to do this this involves the use of **syntactic paths** through trees. Consider again the tree discussed earlier that dominates *American Airlines* and *AMR Inc.* The syntactic relationship between these arguments can be characterized by the path traversed through the tree in getting from one to the other:

$NP \uparrow NP \downarrow NP \downarrow PP \downarrow NP$

Similar path features defined over syntactic dependency trees as well as flat base-phrase chunk structures have been shown to be useful for relation detection and classification (Culotta and Sorensen, 2004; Bunescu and Mooney, 2005). Recall that syntactic path features featured prominently in Ch. 20 in the context of semantic role labeling.

Fig. 22.15 illustrates some of the features that would be extracted while trying to classify the relationship between *American Airlines* and *Tim Wagner* from our example text.

22.2.2 Lightly Supervised Approaches to Relation Analysis

The supervised machine learning approach just described assumes that we have ready access to a large collection of previously annotated material with which to train classifiers. Unfortunately, this assumption is impractical in many real-world settings. A simple approach to extracting relational information without large amounts of annotated material is to use regular expression patterns to match text segments that are likely to contain expressions of the relations in which we are interested.

Consider the problem of building a table containing all the hub cities that various airlines utilize. Assuming we have access a search engine that permits some form of phrasal search with wildcards, we might try something like the following as a query:

```
/ * has a hub at * /
```

Given access to a reasonable amount of material of the right kind, such a search will yield a fair number of correct answers. A recent Google search using this pattern yields the following relevant sentences among the return set.

(22.4) Milwaukee-based Midwest has a hub at KCI.

(22.5) Delta has a hub at LaGuardia.

(22.6) Bulgaria Air has a hub at Sofia Airport, as does Hemus Air.

(22.7) American Airlines has a hub at the San Juan airport.

Of course, patterns such as this can fail in the two ways we discussed all the way back in Ch. 2: by finding some things they shouldn't, and by failing to find things they should. As an example of the first kind of error, consider the following sentences that were also included in the earlier return set.

(22.8) airline j has a hub at airport k

(22.9) The catheter has a hub at the proximal end

(22.10) A star topology often has a hub at its center.

We can address these errors by making our proposed pattern more specific. In this case, replacing the unrestricted wildcard operator with a named entity class restriction would rule these examples out:

```
/[ORG] has a hub at [LOC]/
```

The second problem is that we can't know if we've found all the hubs for all airlines, since we've limited ourselves to this one rather specific pattern. Consider the following close calls missed by our first pattern.

(22.11) No frills rival easyJet, which has established a hub at Liverpool...

(22.12) Ryanair also has a continental hub at Charleroi airport (Belgium).

These examples are missed because they contain minor variations that cause the original pattern to fail. There are two ways to address this problem. The first is to generalize our pattern to capture expressions like these that contain the information we are seeking. This can be accomplished by relaxing the pattern to allow matches that skip parts of the candidate text. Of course, this approach is likely introduce more of the false positives that we tried to eliminate by making our pattern more specific in the first place.

The second, more promising solution, is to expand our set of specific high-precision patterns. Given a large and diverse document collection, an expanded set of patterns should be able to capture more of the information we're looking for. One way to acquire these additional patterns is to simply have human analysts familiar with the domain come up with more patterns and hope to get better coverage. A more interesting automatic alternative is to induce new patterns by **bootstrapping** from the initial search results from a small set of **seed patterns**.

Bootstrapping

Seed patterns

To see how this works, let's assume that we've discovered that Ryanair has a hub at Charleroi. We can use this fact to discover new patterns by finding other mentions of this relation in our corpus. The simplest way to do this is to search for the terms *Ryanair*, *Charleroi* and *hub* in some proximity. The following are among the results from a recent search in Google News.

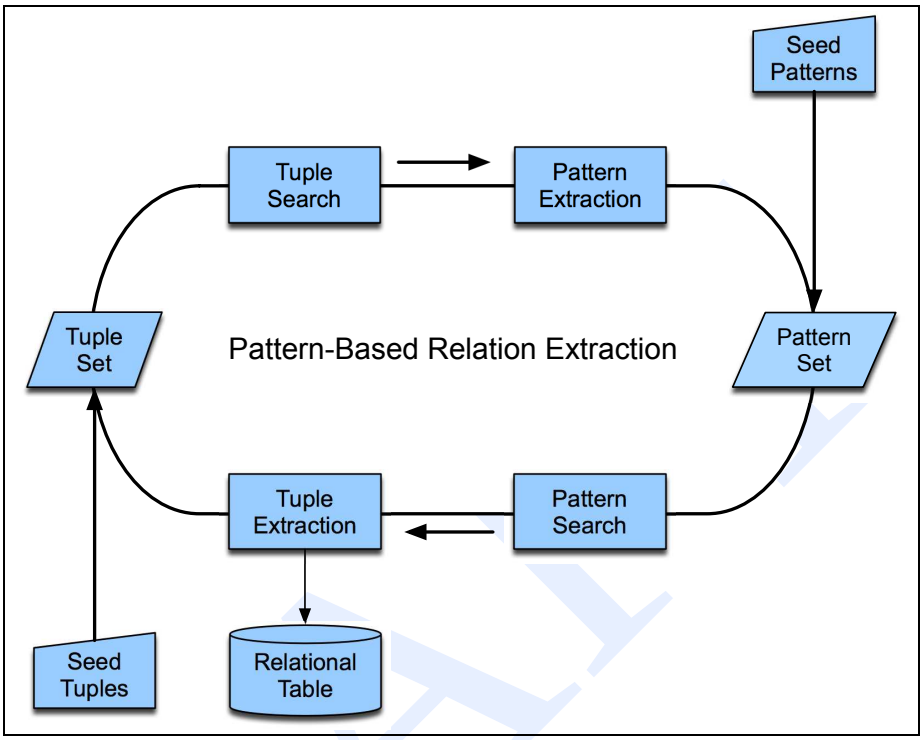

Figure 22.16 Pattern and bootstrapping-based relation extraction.

(22.13) Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.

(22.14) All flights in and out of Ryanair's Belgian hub at Charleroi airport were grounded on Friday...

(22.15) A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had already been affected.

From these results, patterns such as the following can be extracted that look for relevant named entities of various types in the right places.

```
/ [ORG], which uses [LOC] as a hub /
/ [ORG]'s hub at [LOC] /
/ [LOC] a main hub for [ORG] /
```

These new patterns can then be used to search for additional tuples.

Fig. 22.16 illustrates the overall bootstrapping approach. This figure shows that the dual nature of patterns and seeds permits the process to start with either a small set of **seed tuples** or a set of **seed patterns**. This style of bootstrapping and pattern-based relation extraction is closely related to the techniques discussed in Ch. 20 for extracting hyponym and meronym-based lexical relations.

There are, of course, a fair number of technical details to be worked out to actually implement such an approach. The following are among some of the key problems.

- Representing the search patterns

- Assessing the accuracy and coverage of discovered patterns
- And assessing the reliability of the discovered tuples

Patterns are typically represented in a way that captures the following four factors.

- Context prior to the first entity mention
- Context between the entity mentions
- Context following the second mention
- The order of the arguments in the pattern

Contexts are either captured as regular expression patterns or as vectors of features similar to those described earlier for machine learning-based approaches. In either case, they can be defined over character strings, word-level tokens, or syntactic and semantic structures. In general, regular expression approaches tend to be very specific, yielding high precision results; feature-based approaches, on the other hand, are more capable of ignoring potentially inconsequential elements of contexts.

Our next problem is how to assess the reliability of newly discovered patterns and tuples. Recall that we don't, in general, have access to annotated materials giving us the right answers. We therefore have to rely on the accuracy of the initial seed sets of patterns and/or tuples for gold-standard evaluation, and we have to ensure that we don't permit any significant **semantic drift** to occur as we're learning new patterns and tuples. Semantic drift occurs when an erroneous pattern leads to the introduction of erroneous tuples, which can then, turn, lead to the creation of problematic patterns.

Semantic drift

To see this consider the following example.

(22.16) Sydney has a ferry hub at Circular Quay.

If accepted as a positive example, this expression could lead to the introduction of the tuple $\langle Sydney, CircularQuay \rangle$. Patterns based on this tuple could propagate further errors into the database.

There are two factors that need to be balanced in assessing a proposed new pattern: the pattern's performance with respect to the current set of tuples, and the pattern's productivity in terms of the number of matches it produces in the document collection. More formally, given a document collection $\mathscr{D}$, a current set of tuples T, and a proposed pattern p, there are three factors that we need to track.

- $hits$: the set of tuples in T that p matches while looking in $\mathscr{D}$;
- $misses$: The set of tuples in T that p misses while looking at $\mathscr{D}$;
- $finds$: The total set of tuples that p finds in $\mathscr{D}$.

The following equation balances these considerations (Riloff and Jones, 1999).

$$Conf_{RlogF}(p) = \frac{hits_p}{hits_p + misses_p} \times log(finds_p) \tag{22.17}$$

It is useful to be able to treat this metric as a probability, so we'll need to normalize it. A simple way to do this is to track the range of confidences in a development set and divide by some previously observed maximum confidence (Agichtein and Gravano, 2000).

We can assess the confidence in a proposed new tuple by combining the evidence supporting it from all the patterns P' that match that tuple in $\mathscr{D}$ (Agichtein and Gravano,

Noisy-or

2000). One way to combine such evidence is the **noisy-or** technique. Assume that a given tuple is supported by a subset of the patterns in P, each with its own confidence assessed as above. In the noisy-or model, we make two basic assumptions. First, that for a proposed tuple to be false, *all* of its supporting patterns must have been in error, and second that the sources of their individual failures are all independent. If we loosely treat our confidence measures as probabilities, then the probability of any individual pattern p failing is $1 - Conf(p)$; the probability of all of the supporting patterns for a tuple being wrong is the product of their individual failure probabilities, leaving us with the following equation for our confidence in a new tuple.

(22.18)
$$Conf(t) = 1 - \prod_{p \in P'} 1 - Conf(p)$$

The independence assumptions underlying the noisy-or model are very strong indeed. If the failure mode of the patterns are not independent, then the method will overestimate the confidence for the tuple. This overestimate is typically compensated for by setting a very high threshold for the acceptance of new tuples.

Given these measures, we can dynamically assess our confidence in both new tuples and patterns as the bootstrapping process iterates. Setting conservative thresholds for the acceptance of new patterns and tuples should help prevent the system from drifting from the targeted relation.

Although there have been no standardized evaluations for this style of relation extraction on publicly available sources, the technique has gained wide acceptance as a practical way to quickly populate relational tables from open source materials (most commonly from the Web) (Etzioni et al., 2005).

22.2.3 Evaluating Relation Analysis Systems

There are two separate methods for evaluating relation detection systems. In the first approach, the focus is on how well systems can find and classify all the relation mentions in a given text. In this approach, labeled and unlabeled recall, precision and F-measures are used to evaluate systems against a test collection with human annotated gold-standard relations. Labeled precision and recall requires the system to classify the relation correctly, while unlabeled methods simply measure a system's ability to detect entities that are related.

The second approach focuses on the tuples to be extracted from a body of text, rather than on the relation mentions. In this method, systems need not detect every mention of a relation to be scored correctly. Instead, the final evaluation is based on the set of tuples occupying the database when the system is finished. That is, we want to know if the system can discover that RyanAir has a hub at Charleroi; we don't really care how many times it discovers it.

This method has typically used to evaluate unsupervised methods of the kind discussed in the last section. In these evaluations human analysts simply examine the set of tuples produced by the system. Precision is simply the fraction of correct tuples out of all the tuples produced as judged by the human experts.

Recall remains a problem in this approach. It is obviously too costly to search by hand for all the relations that could have been extracted from a potentially large

collection such as the Web. One solution is to compute recall at various levels of precision as described in Ch. 23 (Etzioni et al., 2005). Of course, this isn't true recall, since we're measuring against the number of correct tuples discovered rather than the number of tuples that are theoretically extractable from the text.

Another possibility is to evaluate recall on problems where large resources containing comprehensive lists of correct answers are available. Examples of include gazetteers for facts about locations, the Internet Movie Database (IMDB) for facts about movies or Amazon for facts about books. The problem with this approach is that it measures recall against a database that may be far more comprehensive than the text collections used by relation extraction system.

22.3 Temporal and Event Processing

Our focus thus far has been on extracting information about entities and their relations to one another. However, in most texts, entities are introduced in the course of describing the events in which they take part. Finding and analyzing the events in a text, and how they relate to each other in time, is crucial to extracting a more complete picture of the contents of a text. Such temporal information is particularly important in applications such as question answering and summarization.

In question answering, whether or not a system detects a correct answer may depend on temporal relations extracted from both the question and the potential answer text. As an example of this, consider the following sample question and potential answer text.

When did airlines as a group last raise fares?

> Last week, Delta boosted thousands of fares by $10 per round trip, and most big network rivals immediately matched the increase. (Dateline 7/2/2007).

This snippet does provide an answer to the question, but extracting it requires temporal reasoning to anchor the phrase *last week*, to link that time to the *boosting* event, and finally to link the time of the *matching* event to that.

The following sections introduce approaches to recognizing temporal expressions, figuring out the times that those expressions refer to, detecting events and associating times with those events.

22.3.1 Temporal Expression Recognition

Absolute temporal expressions

Relative temporal expressions

Durations

Temporal expressions are those that refer to absolute points in time, relative times, durations and sets of these. **Absolute temporal expressions** are those that can be mapped directly to calendar dates, times of day, or both. **Relative temporal expressions** map to particular times via some other reference point (as in *a week from last Tuesday*.) Finally, **durations** denote spans of time at varying levels of granularity (seconds, minutes, days, weeks, centuries etc.) Fig. 22.17 provides some sample temporal expressions in each of these categories.

Syntactically, temporal expressions are syntactic constructions that have temporal

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Figure 22.17 Examples of absolute, relation and durational temporal expressions.

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Figure 22.18 Examples of temporal lexical triggers.

Lexical triggers

lexical triggers as their heads. In the annotation scheme in widest use, lexical triggers can be nouns, proper nouns, adjectives, and adverbs; full temporal expression consist of their phrasal projections: noun phrases, adjective phrases and adverbial phrases. Figure 22.18 provides examples of lexical triggers from these categories.

The annotation scheme in widest use is derived from the TIDES standard(Ferro et al., 2005). The approach presented here is based on the TimeML effort (Pustejovsky et al., 2005). TimeML provides an XML tag, TIMEX3, along with various attributes to that tag, for annotating temporal expressions. The following example illustrates the basic use of this scheme (ignoring the additional attributes, which we'll discuss as needed later in Sec. 22.3.2).

> A fare increase initiated <TIMEX3>last week </TIMEX3> by UAL Corp's United Airlines was matched by competitors over <TIMEX3>the weekend </TIMEX3>, marking the second successful fare increase in <TIMEX3>two weeks</TIMEX2>.

The temporal expression recognition task consists of finding the start and end of all of the text spans that correspond to such temporal expressions. Although there are myriad ways to compose time expressions in English, the set of temporal trigger terms is, for all practical purposes, static and the set of constructions used to generate temporal phrases is quite conventionalized. These facts suggest that any of the major approaches to finding and classifying text spans that we've already studied should be successful. The following three approaches have all been successfully employed in recent evaluations.

- Rule-based systems based on partial parsing or chunking
- Statistical sequence classifiers based on standard token-by-token IOB encoding
- Constituent-based classification as used in semantic role labeling

Rule-based approaches to temporal expression recognition use cascades of automata to recognize patterns at increasing levels of complexity. Since temporal expressions are limited to a fixed set of standard syntactic categories, most of these systems make use of pattern-based methods for recognizing syntactic chunks. That is, tokens are first part-of-speech tagged and then larger and larger chunks are recognized using

the results from previous stages. The only difference from the usual partial parsing approaches is the fact that temporal expressions must contain temporal lexical triggers. Patterns must, therefore, contain either specific trigger words (e.g. *February*), or patterns representing classes (e.g. *MONTH*). Fig. 22.19 illustrates this approach with a small representative fragment from a rule-based system written in Perl.

Sequence labeling approaches follow exactly the same scheme introduced in Ch. 13 for syntactic chunking. The three tags I, O and B are used to mark tokens that are either inside, outside or begin a temporal expression, as delimited by TIMEX3 tags. Example 22.3.1 would be labeled as follows in this scheme.

A fare increase initiated last week by UAL Corp's...
O O O O B I O O O

As expected, features are extracted from the context surrounding a token to be tagged and a statistical sequence labeler is trained using those features. As with syntactic chunking and named entity recognition, any of the usual statistical sequence methods can be applied. Fig. 22.20 lists the standard features used in the machine learning-based approach to temporal tagging.

Constituent-based methods combine aspects of both chunking and token-by-token labeling. In this approach, a complete constituent parse is produced by automatic means. The nodes in the resulting tree are then classified, one by one, as to whether they contain a temporal expression or not. This task is accomplished by training a binary classifier with annotated training data, using many of the same features employed in IOB-style training. This approach separates the classification problem from the segmentation problem by assigning the segmentation problem to the syntactic parser. The motivation for this choice was mentioned earlier; in currently available training materials, temporal expressions are limited to syntactic constituents from one of a fixed set of syntactic categories. Therefore, it makes sense to allow a syntactic parser to solve the segmentation part of the problem.

In standard evaluations, temporal expression recognizers are evaluated using the usual recall, precision and F-measures. In recent evaluations, both rule-based and statistical systems achieve about the same level of performance, with the best systems reaching an F-measure of around .87 on a strict exact match criteria. On a looser criterion based on overlap with gold standard temporal expressions, the best systems reach an F-measure of .94.[4]

The major difficulties for all of these approaches are achieving reasonable coverage, correctly identifying the extent of temporal expressions and dealing with expressions that trigger false positives. The problem of false positives arises from the use of temporal trigger words as parts of proper names. For example, all of the following examples are likely to cause false positives for either rule-based or statistical taggers.

(22.19) *1984* tells the story of Winston Smith and his degradation by the totalitarian state in which he lives.

(22.20) Edge is set to join Bono onstage to perform U2's classic *Sunday* Bloody *Sunday*.

[4] http://www.nist.gov/speech/tests/ace/

```
# yesterday/today/tomorrow
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?(($OT+the$CT+\s+)?$OT+day$CT+\s+
$OT+(before|after)$CT+\s+)?$OT+$TERelDayExpr$CT+(\s+$OT+(morning|afternoon|
evening|night)$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gio;

$string =~ s/($OT+\w+$CT+\s+)
<TIMEX2 TYPE=\"DATE\"[^>]*>($OT+(Today|Tonight)$CT+)<\/TIMEX2>/$1$2/gso;

# this/that (morning/afternoon/evening/night)
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?$OT+(this|that|every|the$CT+\s+
$OT+(next|previous|following))$CT+\s*$OT+(morning|afternoon|evening|night)
$CT+(\s+$OT+thereafter$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gosi;
```

Figure 22.19 Fragment of Perl code from MITRE's TempEx temporal tagging system.

Feature	Explanation
Token	The target token to be labeled
Tokens in window	Bag of tokens in the window around a target
Shape	Character shape features
POS	Parts of speech of target and window words
Chunk tags	Base-phrase chunk tag for target and words in a window
Lexical triggers	Presence in a list of temporal terms

Figure 22.20 Typical features used to train IOB style temporal expression taggers.

```
<TIMEX3 id=t1 type="DATE" value="2007-07-02" functionInDocument="CREATION_TIME">
July 2, 2007 </TIMEX3>  A fare increase initiated <TIMEX3 id="t2" type="DATE"
value="2007-W26"  anchorTimeID="t1">last week</TIMEX3> by UAL Corp's United  Airlines
was matched by competitors over <TIMEX3 id="t3" type="DURATION" value="P1WE"
anchorTimeID="t1"> the weekend </TIMEX3>, marking the second successful fare increase
in <TIMEX3 id="t4" type="DURATION"  value="P2W" anchorTimeID="t1"> two weeks </TIMEX3>.
```

Figure 22.21 TimeML markup including normalized values for temporal expressions.

(22.21) Black *September* tried to detonate three car bombs in New York City in March 1973.

22.3.2 Temporal Normalization

Temporal normalization

The task of recognizing temporal expressions is typically followed by the task of normalization. **Temporal normalization** refers to the process of mapping a temporal expression to either a specific point in time, or to a duration. Points in time correspond either to calendar dates or to times of day (or both). Durations primarily consist of lengths of time, but may also include information concerning the start and end points of a duration when that information is available.

Normalized representations of temporal expressions are captured using the VALUE attribute from the ISO 8601 standard for encoding temporal values(ISO8601, 2004). To illustrate some aspects of this scheme, let's return to our earlier example, reproduced in Fig. 22.21 with the value attributes added in.

The dateline, or document date, for this text was *July 2, 2007*. The ISO representation for this kind of expression is YYYY-MM-DD, or in this case, 2007-07-02. The encodings for the temporal expressions in our sample text all follow from this date, and are shown here as values for the VALUE attribute. Let's consider each of these temporal

Unit	Pattern	Sample Value
Fully Specified Dates	YYYY-MM-DD	1991-09-28
Weeks	YYYY-nnW	2007-27W
Weekends	PnWE	P1WE
24 hour clock times	HH:MM:SS	11:13:45
Dates and Times	YYYY-MM-DDTHH:MM:SS	1991-09-28T11:00:00
Financial quarters	Qn	1999-3Q

Figure 22.22 Sample ISO patterns for representing various times and durations.

expressions in turn.

The first temporal expression in the text proper refers to a particular week of the year. In the ISO standard, weeks are numbered from 01 to 53, with the first week of the year being the one that has the first Thursday of the year. These weeks are represented using the template YYYY-Wnn. The ISO week for our document date is week 27, thus the value for *last week* is represented as "2007-W26".

The next temporal expression is *the weekend*. ISO weeks begin on Monday, thus, weekends occur at the end of a week and are fully contained within a single week. Weekends are treated as durations, so the value of the VALUE attribute has to be a length. Durations are represented using the pattern P*nx*, where *n* is an integer denoting the length and *x* represents the unit, as in P3Y for *three years* or P2D for *two days*. In this example, one weekend is captured as P1WE. In this case, there is also sufficient information to anchor this particular weekend as part of a particular week. Such information is encoded in the ANCHORTIMEID attribute. Finally, the phrase *two weeks* also denotes a duration captured as P2W.

There is a lot more to both the ISO 8601 standard and the various temporal annotation standards — far too much to cover here. Fig. 22.22 describes some of the basic ways that other times and durations are represented. Consult (ISO8601, 2004; Ferro et al., 2005; Pustejovsky et al., 2005) for more details.

Most current approaches to temporal normalization employ rule-based methods that associate semantic analysis procedures with patterns matching particular temporal expressions. This is a domain-specific instantiation of the compositional rule-to-rule approach introduced in Ch. 18. In this approach, the meaning of a constituent is computed from the meaning of its parts, and the method used to perform this computation is specific to the constituent being created. The only difference here is that the semantic composition rules involve simple temporal arithmetic rather than λ-calculus attachments.

To normalize temporal expressions, we'll need rules for four kinds of expressions.

- Fully qualified temporal expressions
- Absolute temporal expressions
- Relative temporal expressions
- Durations

Fully qualified date expressions

Fully qualified date expressions contain a year, month and day in some conventional form. The units in the expression must be detected and then placed in the correct place in the corresponding ISO pattern. The following pattern normalizes the fully qualified temporal expression used in expressions like *April 24, 1916*.

$$FQTE \rightarrow Month\ Date\ ,\ Year \qquad \{Year.val - Month.val - Date.val\}$$

In this rule, the non-terminals *Month*, *Date*, and *Year* represent constituents that have already been recognized and assigned semantic values, accessed via the **.val* notation. The value of this *FQE* constituent can, in turn, be accessed as *FQTE.val* during further processing.

Fully qualified temporal expressions are fairly rare in real texts. Most temporal expressions in news articles are incomplete and are only implicitly anchored, often with respect to the dateline of the article, which we'll refer to as the document's **temporal anchor**. The values of relatively simple temporal expressions such as *today*, *yesterday*, or *tomorrow* can all be computed with respect to this temporal anchor. The semantic procedure for *today* simply assigns the anchor, while the attachments for *tomorrow* and *yesterday* add a day and subtract a day from the anchor, respectively. Of course, given the circular nature of our representations for months, weeks, days and times of day, our temporal arithmetic procedures must use modulo arithmetic appropriate to the time unit being used.

Temporal anchor

Unfortunately, even simple expressions such as *the weekend* or *Wednesday* introduce a fair amount of complexity. In our current example, *the weekend* clearly refers to the weekend of the week that immediately precedes the document date. But this won't always be the case, as is illustrated in the following example.

(22.22) Random security checks that began yesterday at Sky Harbor will continue at least through the weekend.

In this case, the expression *the weekend* refers to the weekend of the week that the anchoring date is part of (i.e. the coming weekend). The information that signals this comes from the tense of *continue*, the verb governing *the weekend*.

Relative temporal expressions are handled with temporal arithmetic similar to that used for *today* and *yesterday*. To illustrate this, consider the expression *last week* from our example. From the document date, we can determine that the ISO week for the article is week 27, so *last week* is simply 1 minus the current week.

Again, even simple constructions such as this can be ambiguous in English. The resolution of expressions involving *next* and *last* must take into account the distance from the anchoring date to the nearest unit in question. For example, a phrase such as *next Friday* can refer to either the immediately next Friday, or to the Friday following that. The determining factor has to do with the proximity to the reference time. The closer the document date is to a Friday, the more likely it is that the phrase *next Friday* will skip the nearest one. Such ambiguities are handled by encoding language and domain specific heuristics into the temporal attachments.

The need to associate highly idiosyncratic temporal procedures with particular temporal constructions accounts for the widespread use of of rule-based methods in temporal expression recognition. Even when high performance statistical methods are used for temporal recognition, rule-based patterns are still required for normalization. Although the construction of these patterns can be tedious and filled with exceptions, it appears that sets of patterns that provide good coverage in newswire domains can be created fairly quickly (Ahn et al., 2005).

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character level suffixes for nominalizations (eg. *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
Wordnet hypernyms	Hypernym set for the target

Figure 22.23 Features commonly used in both rule-based and statistical approaches to event detection.

Finally, many temporal expressions are anchored to events mentioned in a text and not directly to other temporal expressions. Consider the following example.

(22.23) One week after the storm, JetBlue issued its customer bill of rights.

To determine when JetBlue issued its customer bill of rights we need to determine the time of *the storm* event, and then that time needs to be modified by the temporal expression *one week after*. We'll return to this issue when we take up event detection in the next section.

22.3.3 Event Detection and Analysis

Event detection and classification

The task of **event detection and classification** is to identify mentions of events in texts and then assign those events to a variety of classes. For the purposes of this task, an event mention is any expression denoting an event or state that can be assigned to a particular point, or interval, in time. The following markup of Example 22.3.1 shows all the events in this text.

> [$_{EVENT}$ Citing] high fuel prices, United Airlines [$_{EVENT}$ said] Friday it has [$_{EVENT}$ increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [$_{EVENT}$ matched] [$_{EVENT}$ the move], spokesman Tim Wagner [$_{EVENT}$ said]. United, a unit of UAL Corp., [$_{EVENT}$ said] [$_{EVENT}$ the increase] took effect Thursday and [$_{EVENT}$ applies] to most routes where it [$_{EVENT}$ competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

In English, most event mentions correspond to verbs, and most verbs introduce events. However, as we can see from our example this is not always the case. Events can be introduced by noun phrases, as in *the move* and *the increase*, and some verbs fail to introduce events, as in the phrasal verb *took effect*, which refers to when the event began rather than to the event itself. Similarly, light verbs such as *make*, *take*, and *have* often fail to denote events. In these cases, the verb is simply providing a syntactic structure for the arguments to an event expressed by the direct object as in *took a flight*.

Both rule-based and statistical machine learning approaches have been applied to the problem of event detection. Both approaches make use of surface information such as parts of speech information, presence of particular lexical items, and verb tense information. Fig. 22.23 illustrates the key features used in current event detection and classification systems.

Having detected both the events and the temporal expressions in a text, the next logical task is to use this information to fit the events into a complete timeline. Such a timeline would be useful for applications such as question answering and summarization. This ambitious task is is the subject of considerable current research but is beyond the capabilities of current systems.

A somewhat simpler, but still useful, task is to impose a partial ordering on the events and temporal expressions mentioned in a text. Such an ordering can provide many of the same benefits as a true timeline. An example of such a partial ordering would be to determine that the fare increase by *American Airlines* came *after* the fare increase by *United* in our sample text. Determining such an ordering can be viewed as a binary relation detection and classification task similar to those described earlier in Sec. 22.2.

Current approaches to this problem attempt to identify a subset of Allen's 13 temporal relations discussed earlier in Ch. 17, and shown here in Fig. 22.24. Recent evaluation efforts have focused on detecting the *before*, *after* and *during* relations among the temporal expressions, document date and event mentions in a text (Verhagen et al., 2007). Most of the top-performing systems employ statistical classifiers, of the kind discussed earlier in Sec. 22.2, trained on the TimeBank corpus (Pustejovsky et al., 2003c).

22.3.4 TimeBank

As we've seen with other tasks, it's tremendously useful to have access to text annotated with the types and relations in which we're interested. Such resources facilitate both corpus-based linguistic research as well as the training of systems to perform automatic tagging. The **TimeBank** corpus consists of text annotated with much of the information we've been discussing throughout this section (Pustejovsky et al., 2003c). The current release (TimeBank 1.2) of the corpus consists of 183 news articles selected from a variety of sources, including the Penn TreeBank and PropBank collections.

TimeBank

Each article in the TimeBank corpus has had the temporal expressions and event mentions in them explicitly annotated in the TimeML annotation (Pustejovsky et al., 2003a). In addition to temporal expressions and events, the TimeML annotation provides temporal links between events and temporal expressions that specify the nature of the relation between them. Consider the following sample sentence and its corresponding markup shown in Fig. 22.25 selected from one of the TimeBank documents.

(22.24) Delta Air Lines soared 33% to a record in the fiscal first quarter, bucking the industry trend toward declining profits.

As annotated, this text includes three events and two temporal expressions. The events are all in the occurrence class and are given unique identifiers for use in further annotations. The temporal expressions include the creation time of the article, which serves as the document time, and a single temporal expression within the text.

In addition to these annotations, TimeBank provides 4 links that capture the temporal relations between the events and times in the text. The following are the within sentence temporal relations annotated for this example.

- Soaring$_{e1}$ is **included** in the fiscal first quarter$_{t58}$

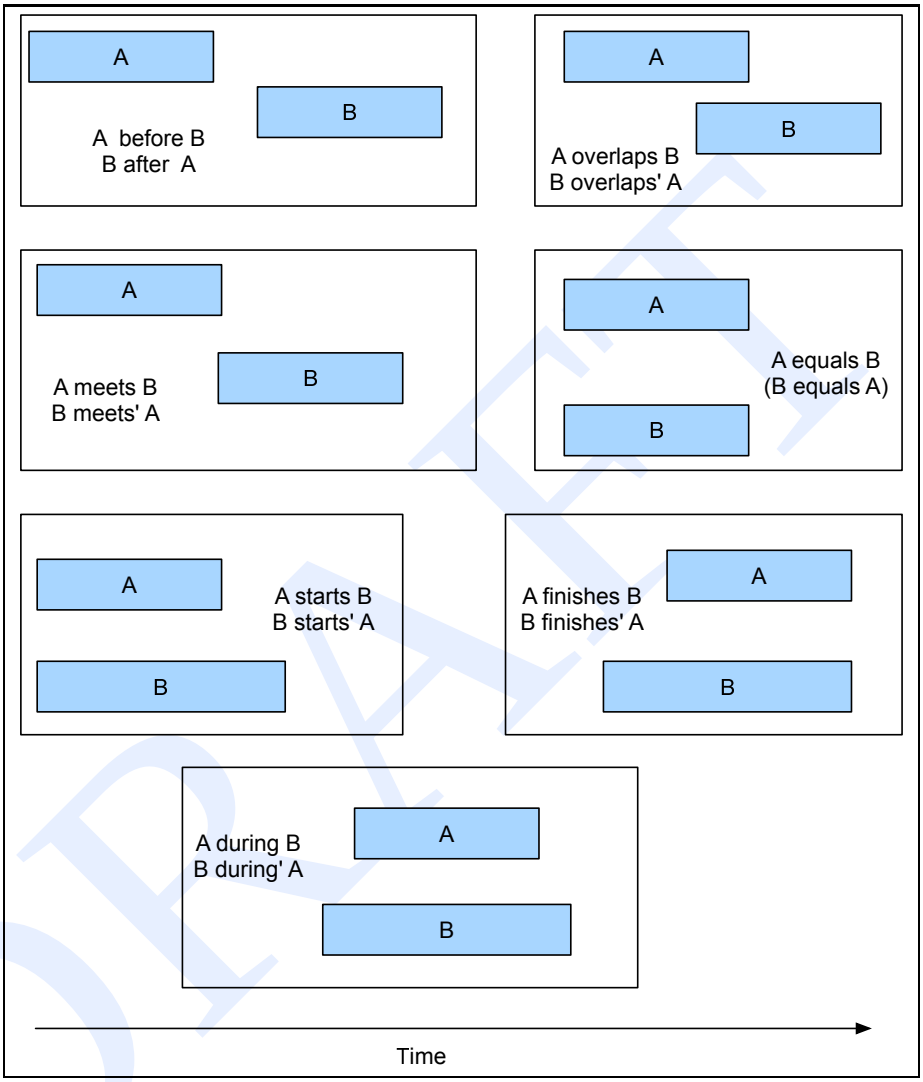

Figure 22.24 Allen's 13 possible temporal relations.

```
<TIMEX3 tid="t57" type="DATE" value="1989-10-26"  functionInDocument="CREATION_TIME">
10/26/89  </TIMEX3>

Delta Air Lines earnings <EVENT eid="e1" class="OCCURRENCE"> soared </EVENT>
33\% to a record in  <TIMEX3 tid="t58" type="DATE" value="1989-Q1" anchorTimeID="t57">
the fiscal first quarter </TIMEX3>, <EVENT eid="e3"  class="OCCURRENCE">bucking</EVENT>
the industry trend toward <EVENT eid="e4" class="OCCURRENCE">declining</EVENT> profits.
```

Figure 22.25 Example from the TimeBank corpus.

- Soaring$_{e1}$ is **before** 1989-10-26$_{t57}$
- Soaring$_{e1}$ is **simultaneous** with the bucking$_{e3}$
- Declining$_{e4}$ **includes** soaring$_{e1}$

The set of 13 temporal relations used in TimeBank are based on Allen's (Allen, 1984) relations introduced earlier in Fig. 22.24.

22.4 Template-Filling

Scripts

Many texts contain reports of events, and possibly sequences of events, that often correspond to fairly common, stereotypical situations in the world. These abstract situations can be characterized as **scripts**, in that they consist of prototypical sequences of sub-events, participants, roles and props (Schank and Abelson, 1977). The use of explicit representations of such scripts in language processing can assist in many of the IE tasks we've been discussing. In particular, the strong expectations provided by these scripts can facilitate the proper classification of entities, the assignment of entities into roles and relations, and most critically, the drawing of inferences that fill in things that have been left unsaid.

Templates

In their simplest form, such scripts can be represented as **templates** consisting of fixed sets of **slots** which take as values **slot-fillers** belonging to particular classes. The task of **template-filling** is to find documents that invoke particular scripts and then fill the slots in the associated templates with fillers extracted from the text. These slot-fillers may consist of text segments extracted directly from the text, or they may consist of concepts that have been inferred from text elements via some additional processing (times, amounts, entities from an ontology, etc.)

A filled template from our original airline story might look like the following.

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Note that as is often the case, the slot-fillers in this example all correspond to detectable named entities of various kinds (organizations, amounts and times). This suggests that template-filling applications should rely on tags provided by named entity recognition, temporal expression and co-reference algorithms to identify candidate slot-fillers.

The next section describes a straightforward approach to filling slots using sequence labeling techniques. Sec. 22.4.2 then describes a system designed to address a considerably more complex template-filling task, based on the use of cascades of finite-state transducers.

22.4.1 Statistical Approaches to Template-Filling

A surprisingly effective approach to template-filling simply casts it as a statistical sequence labeling problem. In this approach, systems are trained to label sequences of tokens as potential fillers for particular slots. There are two basic ways to instantiate this approach: the first is to train separate sequence classifiers for each slot to be filled and then send the entire text through each labeler, the other is to train one large classifier (usually an HMM) that assigns labels for each of the slots to be recognized. We'll focus on the former approach here; we'll take up the single large classifier approach in Ch. 24.

Under the one classifier per slot approach, slots are filled with the text segments identified by each slot's corresponding classifier. As with the other IE tasks described earlier in this chapter, all manner of statistical sequence classifiers have been applied to this problem, all using the usual set of features: tokens, shapes of tokens, part-of-speech tags, syntactic chunk tags, and named entity tags.

There is the possibility in this approach that multiple non-identical text segments will be labeled with the same slot label. This situation can arise in two ways: from competing segments that refer to the same entity using different referring expressions, or from competing segments that represent truly distinct hypotheses. In our sample text, we might expect the segments *United*, *United Airlines* to be labeled as the LEAD AIRLINE. These are not incompatible choices and the reference resolution techniques introduced in Ch. 21 can provide a path to a solution.

Truly competing hypotheses arise when a text contains multiple entities of the expected type for a given slot. In our example, *United Airlines* and *American Airlines* are both airlines and it is possible for both to be tagged as LEAD AIRLINE based on their similarity to exemplars in the training data. In general, most systems simply choose the hypothesis with the highest confidence. Of course, the implementation of this confidence heuristic is dependent on the style of sequence classifier being employed. Markov-based approaches simply select the segment with the highest probability labeling (Freitag and McCallum, 1999).

A variety of annotated collections have been used to evaluate this style of approach to template-filling, including sets of job announcements, conference calls for papers, restaurant guides and biological texts. A frequently employed collection is the CMU Seminar Announcement Corpus[5], a collection of 485 seminar announcements retrieved from the Web with slots annotated for the SPEAKER, LOCATION, START TIME and END TIME. State-of-the-art F-measures on this dataset range from around .98 for the start and end time slots, to as high as .77 for the speaker slot (Roth and tau Yih, 2001; Peshkin and Pfefer, 2003).

As impressive as these results are, they are due as much to the constrained nature of the task as to the techniques they have been employed. Three strong task constraints have contributed to this success. First, in most evaluations all the documents in the collection are all relevant and homogeneous, that is they are known to contain the slots of interest. Second, the documents are all relatively small, providing little room for distractor segments that might incorrectly fill slots. And finally, the target output

[5] http://www.isi.edu/info-agents/RISE/

TIE-UP-1:	
RELATIONSHIP:	TIE-UP
ENTITIES:	"Bridgestone Sports Co."
	"a local concern"
	"a Japanese trading house"
JOINTVENTURECOMPANY	"Bridgestone Sports Taiwan Co."
ACTIVITY	ACTIVITY-1
AMOUNT	NT$20000000
ACTIVITY-1:	
COMPANY	"Bridgestone Sports Taiwan Co."
PRODUCT	"iron and "metal wood" clubs"
STARTDATE	DURING: January 1990

Figure 22.26 The templates produced by the FASTUS (Hobbs et al., 1997) information extraction engine given the input text on page 766.

consists solely of a small set of slots which are to be filled with snippets from the text itself.

22.4.2 Finite-State Template-Filling Systems

The tasks introduced in the *Message Understanding Conferences* (MUC) (Sundheim, 1993), a series of U.S. Government-organized information extraction evaluations, represent a considerably more complex template-filling problem. Consider the following sentences selected from the MUC-5 materials from Grishman and Sundheim (1995).

> Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.
>
> The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

The MUC-5 evaluation task required systems to produce hierarchically linked templates describing the participants in the joint venture, the resulting company, and its intended activity, ownership and capitalization. Fig. 22.26 shows the resulting structure produced by the FASTUS system (Hobbs et al., 1997). Note how the filler of the ACTIVITY slot of the TIE-UP template is itself a template with slots to be filled.

The FASTUS system produces the template given above, based on a cascade of transducers in which each level of linguistic processing extracts some information from the text, which is passed on to the next higher level, as shown in Figure 22.27

Most systems base most of these levels on finite-automata, although in practice most complete systems are not technically finite-state, either because the individual automata are augmented with feature registers (as in FASTUS), or because they are used only as preprocessing steps for full parsers (e.g., Gaizauskas et al., 1995; Weischedel, 1995) , or are combined with other components based on statistical methods (Fisher et al., 1995).

No.	Step	Description
1	Tokens:	Transfer an input stream of characters into a token sequence.
2	Complex Words:	Recognize multi-word phrases, numbers, and proper names.
3	Basic phrases:	Segment sentences into noun groups, verb groups, and particles.
4	Complex phrases:	Identify complex noun groups and complex verb groups.
5	Semantic Patterns:	Identify semantic entities and events and insert into templates.
6	Merging:	Merge references to the same entity or event from different parts of the text.

Figure 22.27 Levels of processing in FASTUS (Hobbs et al., 1997). Each level extracts a specific type of information which is then passed on to the next higher level.

Let's sketch the FASTUS implementation of each of these levels, following Hobbs et al. (1997) and Appelt et al. (1995). After tokenization, the second level recognizes multiwords like *set up*, and *joint venture*, and names like *Bridgestone Sports Co.*. The named entity recognizer is a transducer, composed of a large set of specific mappings designed to handle the usual set of named entities.

The following are typical rules for modeling names of performing organizations like *San Francisco Symphony Orchestra* and *Canadian Opera Company*. While the rules are written using a context-free syntax, there is no recursion and therefore they can be automatically compiled into finite-state transducers.

Performer-Org	→	(pre-location) Performer-Noun+ Perf-Org-Suffix
pre-location	→	locname | nationality
locname	→	city | region
Perf-Org-Suffix	→	orchestra, company
Performer-Noun	→	symphony, opera
nationality	→	Canadian, American, Mexican
city	→	San Francisco, London

The second stage also might transduce sequences like *forty two* into the appropriate numeric value (recall the discussion of this problem in Ch. 8).

The third FASTUS stage implements chunking and produces a sequence of basic syntactic chunks, such as noun groups, verb groups, and so on, using finite-state rules of the sort discussed in Ch. 13.

The output of the FASTUS basic phrase identifier is shown in Figure 22.28; note the use of some domain-specific basic phrases like *Company* and *Location*.

Recall that Ch. 13 described how these basic phrases can be combined into more complex noun groups and verb groups. This is accomplished in Stage 4 of FASTUS, by dealing with conjunction and with the attachment of measure phrases as in the following.

20,000 iron and "metal wood" clubs a month,

and prepositional phrases:

	Phrase Type	Phrase
	Company	Bridgestone Sports Co.
	Verb Group	said
	Noun Group	Friday
	Noun Group	it
	Verb Group	had set up
	Noun Group	a joint venture
	Preposition	in
	Location	Taiwan
	Preposition	with
	Noun Group	a local concern
	Conjunction	and
	Noun Group	a Japanese trading house
	Verb Group	to produce
	Noun Group	golf clubs
	Verb Group	to be shipped
	Preposition	to
	Location	Japan

Figure 22.28 The output of Stage 2 of the FASTUS basic-phrase extractor, which uses finite-state rules of the sort described by Appelt and Israel (1997).

	Template/Slot	Value
1	RELATIONSHIP:	TIE-UP
	ENTITIES:	"Bridgestone Sports Co."
		"a local concern"
		"a Japanese trading house"
2	ACTIVITY:	PRODUCTION
	PRODUCT:	"golf clubs"
3	RELATIONSHIP:	TIE-UP
	JOINTVENTURECOMPANY:	"Bridgestone Sports Taiwan Co."
	AMOUNT:	NT$20000000
4	ACTIVITY:	PRODUCTION
	COMPANY:	"Bridgestone Sports Taiwan Co."
	STARTDATE:	DURING: January 1990
5	ACTIVITY:	PRODUCTION
	PRODUCT:	"iron and "metal wood" clubs"

Figure 22.29 The five partial templates produced by Stage 5 of the FASTUS system. These templates will be merged by the Stage 6 merging algorithm to produce the final template shown in Fig. 22.26 on page 766.

production of 20,000 iron and "metal wood" clubs a month,

The output of Stage 4 is a list of complex noun groups and verb groups. Stage 5 takes this list, ignoring all input that has not been chunked into a complex group, recognizes entities and events in the complex groups, and inserts the recognized objects into the appropriate slots in templates. The recognition of entities and events is done by hand-coded finite-state automata whose transitions are based on particular complex-

phrase types annotated by particular head words or particular features like *company*, *currency*, or *date*.

As an example, the first sentence of the news story above realizes the semantic patterns based on the following two regular expressions (where NG indicates Noun-Group and VG Verb-Group).

- NG(Company/ies) VG(Set-up) NG(Joint-Venture) with NG(Company/ies)
- VG(Produce) NG(Product)

The second sentence realizes the second pattern above as well as the following two patterns:

- NG(Company) VG-Passive(Capitalized) at NG(Currency)
- NG(Company) VG(Start) NG(Activity) in/on NG(Date)

The result of processing these two sentences is the set of five draft templates shown in Fig. 22.29. These five templates must then be merged into the single hierarchical structure shown in Fig. 22.26. The merging algorithm decides whether two activity or relationship structures are sufficiently consistent that they might be describing the same events, and merges them if so. The merging algorithm must also perform reference resolution as described in Ch. 21.

22.5 Advanced: Biomedical Information Extraction*

Information extraction from biomedical journal articles has become an important application area in recent years. The motivation for this work comes primarily from biologists, who find themselves faced with an enormous increase in the number of publications in their field since the advent of modern genomics — so many that keeping up with the relevant literature is nearly impossible for many scientists. Fig. 22.30 amply demonstrates the severity of the problem faced by these scientists. Clearly, applications that can automate the extraction and aggregation of useful information from such sources would be a boon to researchers.

A growing application area for information extraction in the biomedical domain is as an aid to the construction of large databases of genomic and related information. Without the availability of information extraction-based curator assistance tools, many manual database construction efforts will not be complete for decades — a time-span much too long to be useful (Baumgartner et al., 2007).

A good example of this kind of application is the MuteXt system. This system targets two named entity types — mutations in proteins and two very specific types of proteins called *G-coupled protein receptors and nuclear hormone receptors*. MuteXt was used to build a database that drew information from 2,008 documents; building it by hand would have taken an enormously time-consuming and expensive undertaking. Mutations in G-protein coupled receptors are associated with a range of diseases that includes diabetes, ocular albinism, and retinitis pigmentosa, so even this simple text mining system has a clear application to the relief of human suffering.

*This section was written by Kevin Bretonnel Cohen

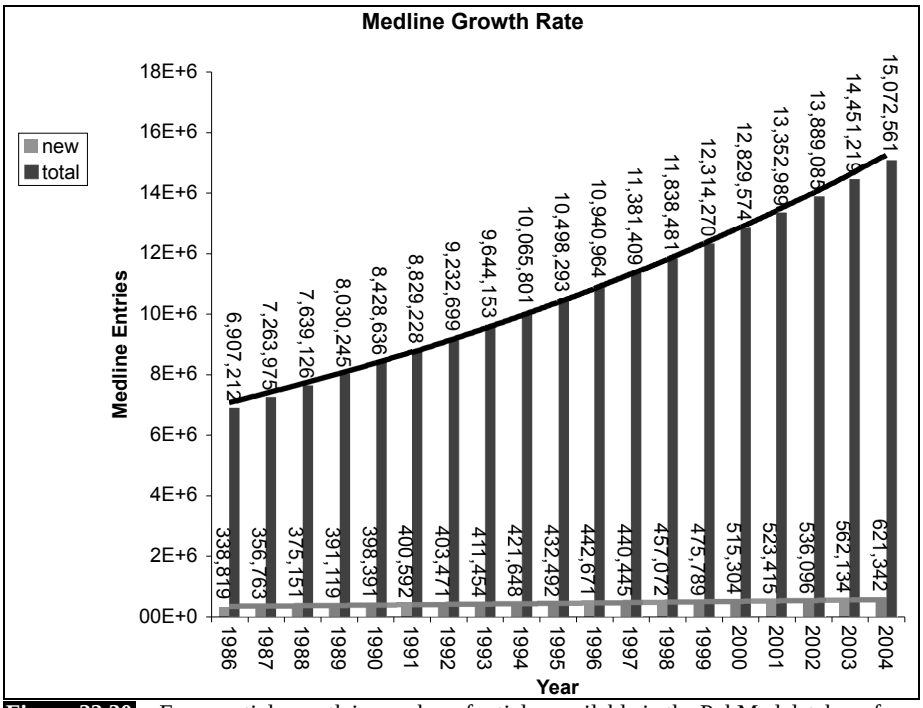

Figure 22.30 Exponential growth in number of articles available in the PubMed database from 1986 to 2004 (Data from (Cohen and Hunter, 2004)).

Biologists and bioinformaticians have recently come up with even more innovative uses for text mining systems, in which the output is never intended for viewing by humans, but rather is used as part of the analysis of high-throughput assays—experimental methods which produce masses of data points that would have been unimaginable just twenty years ago—and as part of techniques for using data in genomic data repositories. Ng (2006) provides a review and an insightful analysis of work in this vein.

22.5.1 Biological Named Entity Recognition

Information extraction tasks in the biological realm are characterized by a much wider range of relevant types of entities than the PERSON, ORGANIZATION, and LOCATION semantic classes that characterize work that is focused on news-style texts. Fig. 22.31 and the following example illustrate just a small subset of the variety of semantic classes of named entities that have been the target of NER systems in the biomedical domain.

> [$_{TISSUE}$ Plasma] [$_{GP}$ BNP] concentrations were higher in both the [$_{POPULATION}$ judo] and [$_{POPULATION}$ marathon groups] than in [$_{POPULATION}$ controls], and positively correlated with [$_{ANAT}$ LV] mass as well as with deceleration time.

Nearly all of the techniques described in Sec. 22.1 have been applied to the biomedical NER problem, with a particular focus on the problem of recognizing gene/protein

Semantic class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64 → Gly*
Populations	*judo group*

Figure 22.31 A sample of the semantic classes of named entities that have been recognized in biomedical NLP. Note the surface similarities between many of the examples.

names. This task is particularly difficult due to the wide range of forms that gene names can take: *white*, *insulin*, *BRCA1*, *ether a go-go*, and *breast cancer associated 1* are all the names of genes. The choice of algorithm for gene name recognition seems to be less important than the choice of features; typical feature sets include word-shape and contextual features, as discussed earlier; additionally, knowledge-based features, such as using the count of Google hits for a sequence like *BRCA1 gene* to decide whether or not a token of the string *BRCA1* is a reference to a gene or not, are sometimes incorporated into statistical systems.

Surprisingly, the use of huge publicly available lists of gene names has not generally contributed to the performance of a gene/protein NER system (Yeh et al., 2005), and in fact may actually degrade it (Baumgartner et al., 2006). It is not uncommon for gene names to be many tokens long (e.g. *breast cancer associated 1*). Gene name length has a demonstrable effect on NER system performance (Kinoshita et al., 2005; Yeh et al., 2005), and any technique for correctly finding the boundaries of multi-token names seems to increase performance. Use of the abbreviation-definition-detection algorithm (Schwartz and Hearst, 2003) is common for this purpose, since many such names appear as abbreviation or symbol definitions at some point in a publication. Base noun group chunkers can also be useful in this regard, as can a surprisingly small number of heuristic rules (Kinoshita et al., 2005).

22.5.2 Gene Normalization

Gene normalization

Having identified all the mentions of biological entities in a text, the next step is to map them to unique identifiers in databases or ontologies. This task has been most heavily studied for genes, where it is known as **gene normalization**. Some of the complexities of the problem come from high degrees of variability in the realization of the names of specific entities in naturally-occurring text; the nature of the problem was first delineated by Cohen et al. (2002). In that work a standard discovery procedure from descriptive linguistics was used to determine what sorts of variability in gene names can be ignored, and what sorts must not be ignored. More recently, Morgan et al. (2007) have shown how linguistic characteristics of community-specific gene-naming conventions affect the complexity of this task when the normalization of genes from

varying species is attempted. Gene normalization can be considered a type of word sense disambiguation task, midway between a targetted WSD task and an all-words WSD task.

An important thread of work on this problem involves mapping named entities to biomedical ontologies, especially the Gene Ontology (Ashburner et al., 2000). This has proven considerably more challenging; terms in the Gene Ontology tend to be long, to have many possible lexical and syntactic forms, and to sometimes require significant amounts of inference.

22.5.3 Biological Roles and Relations

Finding and normalizing all the mentions of biological entities in a text is a preliminary step to determining the roles played by entities in the text. Two ways to do this that have been the focus of recent research are to discover and classify the expressed binary relations between the entities in a text, and to identify and classify the roles played by entities with respect to the central events in the text. These two tasks correspond roughly to the tasks of classifying the relationship between pairs of entities as described in Sec. 22.2, and to the semantic role labeling task introduced in Ch. 20.

Consider the following example texts that express binary relations between entities.

(22.25) These results suggest that con A-induced [$_{DISEASE}$ hepatitis] was ameliorated by pretreatment with [$_{TREATMENT}$ TJ-135].

(22.26) [$_{DISEASE}$ Malignant mesodermal mixed tumor of the uterus] following [$_{TREATMENT}$ irradiation]

Each of these examples asserts a relationship between a *disease* and a *treatment*. In the first example, the relationship can be classified as that of *curing*. In the second example, the disease is a *result* of the mentioned treatment. Rosario and Hearst (2004) present a system for the classification of 7 kinds disease-treatment relations. In this work, a series of HMM-based generative models as well as a discriminative neural network model were successfully applied.

More generally, a wide-range of rule-based and statistical approaches have been applied to binary relation recognition problems such as this. Examples of other widely studied biomedical relation recognition problems include genes and their biological functions (Blaschke et al., 2005), genes and drugs (Rindflesch et al., 2000), genes and mutations (Rebholz-Schuhmann et al., 2004), and protein-protein interactions (Rosario and Hearst, 2005).

Now consider the following example that corresponds to a semantic role labeling style of problem.

(22.27) [$_{THEME}$ Full-length cPLA2] was [$_{TARGET}$ phosphorylated] stoichiometrically by [$_{AGENT}$ p42 mitogen-activated protein (MAP) kinase] in vitro... and the major site of phosphorylation was identified by amino acid sequencing as [$_{SITE}$ Ser505]

The *phosphorylation* event that lies at the core of this text has three semantic roles associated with it: the causal AGENT of the event, the THEME or entity being phosphorylated and finally the location, or SITE of the event. The problem is to identify the

constituents in the input that play these roles and assign them the correct role labels. Note that this example, contains a further complication in that the second event mention *phosphorylation* must be identified as coreferring with the first *phosphorylated* in order to capture the SITE role correctly.

Much of the difficulty with semantic role labeling in the biomedical domain stems from the preponderance of nominalizations in these texts. Nominalizations like *phosphorylation* typically offer fewer syntactic cues to signal their arguments than their verbal equivalents, making the identification task more difficult. A further complication is that different semantic roles arguments often occur as parts of the same, or dominating nominal constituents. To see this consider the following examples.

(22.28) Serum stimulation of fibroblasts in floating matrices does not result in [$_{TARGET}$ [$_{ARG1}$ ERK] translocation] to the [$_{ARG3}$ nucleus] and there was decreased serum activation of upstream members of the ERK signaling pathway, MEK and Raf,

(22.29) The translocation of RelA/p65 was investigated using Western blotting and immunocytochemistry. the COX-2 inhibitor SC236 worked directly through suppressing [$_{TARGET}$ [$_{ARG3}$ nuclear] translocation] of [$_{ARG1}$ RelA/p65].

(22.30) Following UV treatment, Mcl-1 protein synthesis is blocked, the existing pool of Mcl-1 protein is rapidly degraded by the proteasome, and [$_{ARG1}$ [$_{ARG2}$ cytosolic] Bcl-xL] [$_{TARGET}$ translocates] to the [$_{ARG3}$ mitochondria]

Each these examples contains arguments that are bundled into constituents with other arguments or with the target predicate itself. For example, in the second example the constituent *nuclear translocation* signals both the TARGET and the ARG3 role.

Both rule-based and statistical approaches have been applied to these semantic role-like problems. As with relation-finding and NER, the choice of algorithm is less important than the choice of features, many of which are derived from accurate syntactic analyses. However, since there are no large treebanks available for biological texts, we are left with the option using off-the-shelf parsers trained on generic newswire texts. Of course, the errors introduced in this process may negate whatever power we can derive from syntactic features. Therefore, an important area of research revolves around the adaptation of generic syntactic tools to this domain (Blitzer et al., 2006).

Relational and event extraction applications in this domain often have an extremely limited foci. The motivation for this is that even systems with narrow scope can make a contribution to the productivity of working bioscientists. An extreme example of this is the RLIMS-P system discussed earlier. It tackles only the verb *phosphorylate* and the associated nominalization *phosphorylization*. Nevertheless, this system was successfully used to produce a large online database that is in widespread use by the research community.

As the targets of biomedical information extraction applications have become more ambitious, the range of BioNLP application types has become correspondingly more broad. Computational lexical semantics and semantic role labelling (Verspoor et al., 2003; Wattarujeekrit et al., 2004; Ogren et al., 2004; Kogan et al., 2005; Cohen and Hunter, 2006), summarization (Lu et al., 2006), and question-answering are all active research topics in the biomedical domain. Shared tasks like BioCreative continue to be a source of large data sets for named entity recognition, question-answering, relation

extraction, and document classification (Hirschman and Blaschke, 2006), as well as a venue for head-to-head assessment of the benefits of various approaches to information extraction tasks.

22.6 Summary

This chapter has explored a series of techniques for extracting limited forms of semantic content from texts. Most techniques can be characterized as problems in detection followed by classification.

- **Named entities** can be recognized and classified by **statistical sequence labeling** techniques.
- **Relations among entities** can be detected and classified using supervised learning methods when annotated training data is available; lightly supervised **bootstrapping** methods can be used when small numbers of **seed tuples** or **seed patterns** are available.
- Reasoning about time can be facilitated by detecting and normalizing **temporal expressions** through a combination of statistical learning and rule-based methods.
- Rule-based and statistical methods can be used to detect, classify and order **events** in time. The **TimeBank corpus** can facilitate the training and evaluation of temporal analysis systems.
- **Template-filling** applications can recognize stereotypical situations in texts and assign elements from the text to roles represented as **fixed sets of slots**.
- Information extraction techniques have proven to be particularly effective in processing texts from the **biological domain**.
- Scripts, plans and goals...

Bibliographical and Historical Notes

The earliest work on information extraction addressed the template-filling task and was performed in the context of the Frump system (DeJong, 1982a). Later work was stimulated by the U.S. government sponsored MUC conferences (Sundheim, 1991, 1992, 1993, 1995b). Chinchor et al. (1993) describes the evaluation techniques used in the MUC-3 and MUC-4 conferences. Hobbs (1997) partially credits the inspiration for FASTUS to the success of the University of Massachusetts CIRCUS system (Lehnert et al., 1991) in MUC-3. The SCISOR system is another system based loosely on cascades and semantic expectations that did well in MUC-3 (Jacobs and Rau, 1990).

Due to the difficulty of reusing or porting systems from one domain to another, attention shifted to the problem of automatic knowledge acquisition for these systems. The earliest supervised learning approaches to IE are described in Cardie (1993),

Cardie (1994), Riloff (1993), Soderland et al. (1995), Huffman (1996), and Freitag (1998).

These early learning efforts focused on automating the knowledge acquisition process for mostly finite-state rule-based systems. Their success, and the earlier success of HMM-based methods for automatic speech recognition, led to the development of statistical systems based on sequence labeling. Early efforts applying HMMs to IE problems include the work of Bikel et al. (1997, 1999) and Freitag and McCallum (1999). Subsequent efforts demonstrated the effectiveness of a range of statistical methods including MEMMs (McCallum et al., 2000), CRFs (Lafferty et al., 2001) and SVMs (Sassano and Utsuro, 2000; McNamee and Mayfield, 2002).

Progress in this area continues to be stimulated by formal evaluations with shared benchmark datasets. The MUC evaluations of the mid-1990s were succeeded by the Automatic Content Extraction (ACE) program evaluations held periodically from 2000 to 2007.[6] These evaluations focused on the named entity recognition, relation detection, and temporal expression detection and normalization tasks. Other IE evaluations include the 2002 and 2003 CoNLL shared tasks on language-independent named entity recognition (Sang, 2002; Sang and De Meulder, 2003), and the 2007 SemEval tasks on temporal analysis (Verhagen et al., 2007) and people search (Artiles et al., 2007).

The scope of information extraction continues to expand to meet the ever-increasing needs of applications for novel kinds of information. Some of the emerging IE tasks that we haven't discussed include the classification of gender (Koppel et al., 2002), moods (Mishne and de Rijke, 2006), sentiment, affect and opinions (Qu et al., 2004). Much of this work involves **user generated content** in the context of **social media** such as blogs, discussion forums, newsgroups and the like. Research results in this domain have been the focus of a number of recent workshops and conferences (Nicolov et al., 2006; Nicolov and Glance, 2007).

User generated content
Social media

Exercises

22.1 Develop a set of regular expressions to recognize the character shape features described in Fig. 22.7.

22.2 Using a statistical sequence modeling toolkit of your choosing, develop and evaluate an NER system.

22.3 The IOB labeling scheme given in this chapter isn't the only possible one. For example, an E tag might be added to mark the end of entities, or the B tag can be reserved only for those situations where an ambiguity exists between adjacent entities. Propose a new set of IOB tags for use with your NER system. Perform experiments and compare its performance against the scheme presented in this chapter.

22.4 Names of works of art (books, movies, video games, etc.) are quite different from the kinds of named entities we've discussed in this chapter. Collect a list of

[6] www.nist.gov/speech/tests/ace/

names of works of art from a particular category from a web-based source (eg. gutenberg.org, amazon.com, imdb.com, etc.). Analyze your list and give examples of ways that the names in it are likely to be problematic for the techniques described in this chapter.

22.5 Develop an NER system specific to the category of names that you collected in the last exercise. Evaluate your system on a collection of text likely to contain instances of these named entities.

22.6 Acronym expansion, the process of associating a phrase with a particular acronym, can be accomplished by a simple form of relational analysis. Develop a system based on the relation analysis approaches described in this chapter to populate a database of acronym expansions. If you focus on English **Three Letter Acronyms** (TLAs) you can evaluate your system's performance by comparing it to Wikipedia's TLA page (`en.wikipedia.org/wiki/Category:Lists_of_TLAs`).

22.7 Collect a corpus of biographical Wikipedia entries of prominent people from some coherent area of interest (sports, business, computer science, linguistics, etc.). Develop a system that can extract an occupational timeline for the subjects of these articles. For example, the Wikipedia entry for Peter Norvig might result in the ordering: Sun, Harlequin, Junglee, NASA, Google; the entry for David Beckham would be: Manchester United, Real Madrid, Los Angeles Galaxy.

22.8 A useful functionality in newer email and calendar applications is the ability to associate temporal expressions associated with events in emails (doctor's appointments, meeting planning, party invitations, etc.) with specific calendar entries. Collect a corpus of emails containing temporal expressions related to event planning. How do these expressions compare to the kind of expressions commonly found in news text that we've been discussing in this chapter?

22.9 Develop and evaluate a recognition system capable of recognizing temporal expressions of the kind appearing in your email corpus.

22.10 Design a system capable of normalizing these expressions to the degree required to insert them into a standard calendaring application.

22.11 Acquire the CMU seminar announcement corpus and develop a template-filling system using any of the techniques mentioned in Sec. 22.4. Analyze how well your system performs as compared to state-of-the-art results on this corpus.

22.12 Develop a new template that covers a situation commonly reported on by standard news sources. Carefully characterize your slots in terms of the kinds of entities that appear as slot-fillers. Your first step in this exercise should be to acquire a reasonably sized corpus of stories that instantiate your template.

22.13 Given your corpus, develop an approach to annotating the relevant slots in your corpus so that it can serve as a training corpus. Your approach should involve some hand-annotation, but should not be based solely on it.

22.14 Retrain your system and analyze how well it functions on your new domain.

22.15 Species identification is a critical issue for biomedical information extraction applications such as document routing and classification. But it is especially crucial for realistic versions of the gene normalization problem.

Build a species identification system that works on the document level, using the machine learning or rule-based method of your choice. As gold standard data, use the BioCreative gene normalization data (`biocreative.sourceforge.net`).

22.16 Build, or borrow, a named entity recognition system that targets mentions of genes and gene products in texts. As development data, use the BioCreative gene mention corpus (`biocreative.sourceforge.net`).

22.17 Build a gene normalization system that maps the output of your gene mention recognition system to the appropriate database entry. Use the BioCreative gene normalization data as your development and test data, *be sure you don't give your system access to the species identification in the metadata.*