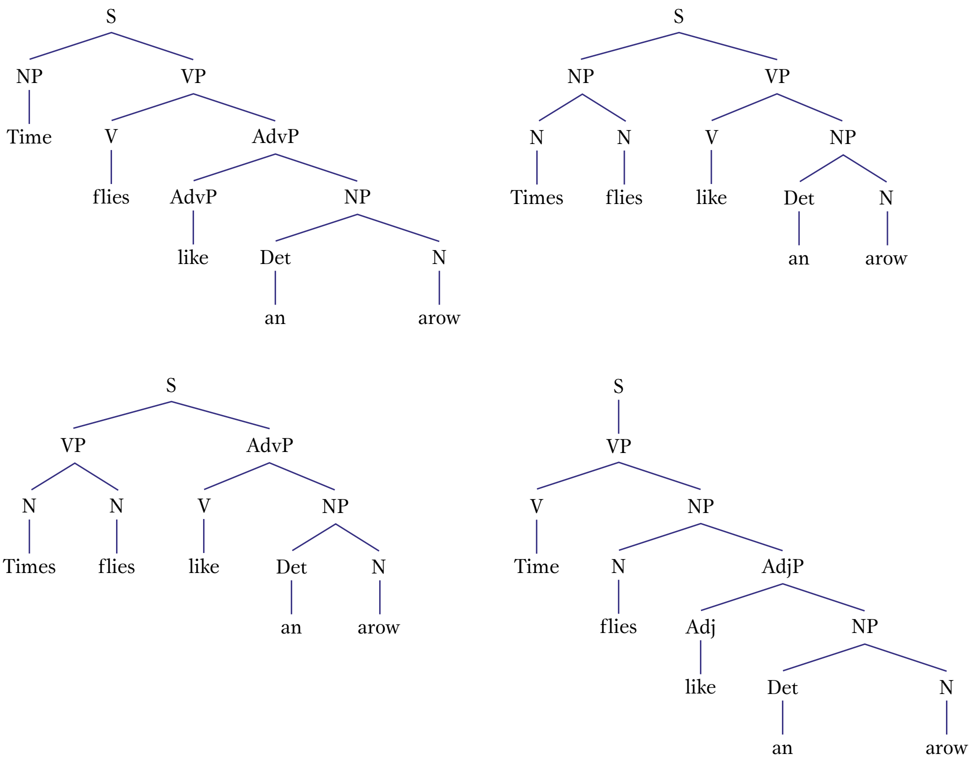
6•12 Statistical parsing

Malouf R.P., Nerbonne J.

A fundamental problem facing developers of natural language processing (NLP) systems is that the grammatical constraints created by linguists admit structures in language which no human would recognize. For example, a sentence like “The tourist saw museums” sounds simple enough, but most NLP systems will recognize not only the intended meaning, but also the meaning in which “saw” is a noun, and the entire string is parsed as a determiner “the” followed by a compound noun “tourist saw museums”. This reading is nonsensical, but cannot be ruled out on purely structural grounds without also ruling out the parallel structure in “the circular saw blades”. Human speakers, on the other hand, have no problem with this kind of ambiguity. Clearly, the purely structural view of language is missing crucial aspects of how language is used in everyday life.

The central importance of disambiguation, or of finding the intended reading among the many readings produced by a parser, has been recognized at least since 1963, when a Harvard University research team headed by Susumo Kuno tested their parser with the sentence “Time flies like an arrow” and were rewarded with four separate readings (see figure 18). A typical architecture for disambiguation uses a probabilistic context free rule system, where estimates of rule probabilities are derived from the frequency with which rules have been encountered in collections of parses, which have been disambiguated by hand. With a sufficient quantity of annotated training data and careful selection of stochastic features, such systems perform adequately enough on structural disambiguation tasks to support simple applications.

Figure 18 ‣ Four readings of one sentence



More sophisticated applications such as open-domain question answering or dialog systems, however, require more sophisticated grammar formalism, which come closer to the richness of language as used and processed by human speakers. And, since these formalisms involve a more complex flow of information than simple context-free grammars, more complex statistical methods are required to capture the subtle dependencies among grammatical structures. Current research on disambiguation is focused on the development of more sophisticated statistical models which allow integration of syntactic, semantic, and discourse information, while at the same time reducing the requirement for large quantities of hand-annotated data.