Exploiting Statistical Information in Constructional Analysis

John Bryant
International Computer Science Institute

Recent work in linguistics argues that statistical information should be considered a part of grammar [2]. Consistent with this hypothesis, statistical information plays an important role in the process of mapping an utterance onto its meaning using constructions. We call this process Constructional Analysis. The implemented analysis system takes as input an utterance and a grammar specified in the Embodied Construction Grammar formalism [1]. It searches for the most likely analysis of the utterance given the grammar’s syntactic and semantic constraints. The system then outputs both the constructions used to interpret the utterance and the utterance’s semantic representation in terms of frames and image schemas.

The program’s search process can also be used to predict which words in an utterance are associated with increased reading time, but the accuracy of the predictions depends crucially on the inclusion of the following probabilistic parameters:

1. $\Pr(\text{omitted, local} \mid \alpha, \beta)$: The probability of construction $\alpha$’s constituent $\beta$ being expressed locally (as a normal constituent) or omitted altogether. This term represents a construction’s preference for omitting each of its constituents.

2. $\Pr(\text{filler} = \theta \mid \alpha, \beta, \text{local}(\beta))$: Assuming construction $\alpha$’s constituent $\beta$ is expressed locally, this term specifies the likelihood of a particular constructional filler $\theta$. For example, this term encodes the subject-predicate construction’s preferences on the kinds of noun phrases that fill its subject role.

3. $\Pr(\text{frameRole} = \rho \mid \gamma, \sigma)$: This parameter encodes the likelihood of a particular semantic filler $\sigma$ being assigned role $\rho$ in frame $\gamma$. e.g. The difference between a criminal being assigned the agent role versus the patient role in the Arrest frame.

Using these parameters, the model generates reading time predictions that follow the trends found by McRae, Spivey and Tannenhaus [3] in their experiments with reduced relative sentences such as *The cop arrested by the detective was guilty*. The fact that these predictions can only be made if the grammar is augmented with probabilistic parameters provides further evidence that statistical information should be considered a part of grammar.

References

