This paper discusses a maximum entropy approach for natural language processing, which uses features and constraints as building blocks and obeys the maximum entropy principle. The principle chooses the model with greatest entropy from all models consistent with constraints. And it yields the same result with another philosophy maximum likelihood. To construct models, (1) the selection of a set of features and (2) computation of parameters of a model with included features are needed. In addition, a greedy method is introduced to make costly computation feasible. Finally, several applications of the model are explored, which include bilingual sense disambiguation, word reordering and sentence segmentation.

In the paper, the general research object is a random process that produces an output value y influenced by contextual information x, and the task is to build a stochastic model representing the behavior of the process, or the probability distribution p(y|x). Then 2 important and relevant concepts are introduced: feature (or feature function) and constraint (or constraint equation). A feature is a binary-valued function, while a constraint is an equation between the expected value of the feature in the *model* and its expected value in the *training data*. The maximum entropy philosophy “models all that is known and assume nothing about that which is unknown”, which indicates that a distribution with maximum conditional entropy gives best “uniformity”. Such a constrained optimization problem to find the model maximizes entropy from a set of allowed probability distribution is called the primal problem.

Moreover, there exists a dual problem (as an unconstrained optimization problem) to find the “parameters” (Lagrange multipliers) maximizing the corresponding dual function. According to the fundamental principle in the theory of Lagrange multipliers, Kuhn-Tucker theorem, the model with *maximum entropy* in the primal problem *is* the model in the parametric family (the dual problem) that *maximizes the likelihood* of the training data. To compute the parameters (dual problem, and then find solution for the primal problem), an algorithm using improved iterative scaling is discussed, for which the key step is the computation of the increments of Lagrange multipliers.

There is one more important problem to address: finding appropriate facts about the data, or feature selection. The authors introduce a strategy to start with a large model space and then incrementally shrink it by adding more features. The gain as the difference of training data log-likelihood between the best allowed models with/without a new feature is used to measure the “influence” of including this new feature. During each iteration, the feature with greatest gain in the set of remaining candidates is chosen to add into current set of features. Notice that the abovementioned algorithm to solve the dual problem is called to find the best parameter/model for each candidate feature. Furthermore, to avoid the expensive computation for the gain, a greedy method assumes that the optimal parameter values of already-included features are unchanged, and thus what is only needed is doing line search for the parameter value of the candidate feature, which minimizes the “approximate gain”.