model is independent of the keyword for disambiguation. Our system does not need human-annotated instances for each target ambiguous word. The weak supervision is performed by using a limited amount of existing annotated corpus which does not need to include the target word set.

The insight is that the correlation regularity between the sense distinction and the context distinction can be captured at Part-of-Speech category level, independent of individual words or word senses. Since context determines the sense of a word, a reasonable hypothesis is that there is some mechanism in the human comprehension process that will decide when two contexts are similar (or dissimilar) enough to trigger our interpretation of a word in the contexts as one meaning (or as two different meanings). We can model this mechanism by capturing the sense distinction regularity at category level.

In the light of this, a maximum entropy model is trained to determine if a pair of contexts of the same keyword refers to the same or different word senses. The maximum entropy modeling is based on heterogeneous context features that involve both trigger words and parsing structures. To ensure the resulting model's independency of individual words, the keywords used in training are different from the keywords used in benchmarking. For any target keyword, a collection of contexts is retrieved from a large raw document pool. Context clustering is performed to derive the optimal context clusters which globally fit the local context pair classification results. Here statistical annealing is used for its optimal performance. In benchmarking, a mapping procedure is required to correlate the context clusters with external ontology senses.

In what follows, Section 2 formulates the maximum entropy model for context pair classification. The context clustering algorithm, including the object function of the clustering and the statistical annealing-based optimization, is described in Section 3. Section 4 presents and discusses benchmarks, followed by conclusion in Section 5.

2 Maximum Entropy Modeling for Context Pair Classification

Given n mentions of a keyword, we first introduce the following symbols. C_i refers to the i-th context. S_i refers to the sense of the i-th context. $CS_{i,j}$ refers to the context similarity between the i-th context and the j-th context, which is a subset of the predefined context similarity features. f_α refers to the α-th predefined context similarity feature. So $CS_{i,j}$ takes the form of $\{f_\alpha\}$.

In this section, we study the context pair classification task, i.e. given a pair of contexts C_i and C_j of the same target word, are they referring to the same sense? This task is formulated as comparing the following conditional probabilities: $\Pr(S_i = S_j | CS_{i,j})$ and $\Pr(S_i \neq S_j | CS_{i,j})$. Unlike traditional context classification for WSD where statistical model is trained for each individual word, our context pair classification model is trained for each Part-of-speech (POS) category. The reason for choosing POS as the appropriate category for learning the context similarity is that the parsing structures, hence the context representation, are different for different POS categories.

The training corpora are constructed using the Senseval-2 English Lexical Sample training corpus. To ensure the resulting model's independency of individual words, the target words used for benchmarking (which will be the ambiguous words used in Senseval-3 English Lexicon Sample task) are carefully removed in the corpus construction process. For each POS category, positive and negative instances are constructed as follows.

Positive instances are constructed using context pairs referring to the same sense of a word. Negative instances are constructed using context pairs that refer to different senses of a word.

For each POS category, we have constructed about 36,000 instances, half positive and half negative. The instances are represented as pairwise context similarities, taking the form of $\{f_\alpha\}$.

Before presenting the context similarity features we used, we first introduce the two categories of the involved context features:

i) Co-occurring trigger words within a predefined window size equal to 50 words to both sides of the keyword. The trigger words are learned from a TIPSTER document pool containing ~170 million words of AP and WSJ news articles. Following (Schütze 1998), χ^2 is used to measure the cohesion between the keyword and a co-occurring word. In our ex-