

**Diagnosing and Minimizing Semantic Drift in Iterative Bootstrapping Extraction**

**Abstract:**

Semantic drift is a common problemin iterative information extraction. Previous approaches for minimizing semantic drift may incur substantial loss in recall. We observe that most semantic drifts are introduced by a small number of questionable extractions in the earlier rounds of iterations. These extractions subsequently introduce a large number of questionable results, which lead to the semantic drift phenomenon.We call these questionable extractions Drifting Points (DPs). If erroneous extractions are the “symptoms” of semantic drift, then DPs are the “causes” of semantic drift. In this paper, we propose a method to minimize semantic drift by identifying the DPs and removing the effect introduced by the DPs. We use isA (concept-instance) extraction as an example to describe our approach in cleaning information extraction errors caused by semantic drift, but we performexperiments on different relation extraction processes on three large real data extraction collections. The experimental results show that our DP cleaning method enables us to clean around 90 percent incorrect instances or patterns with about 90 percent precision, which outperforms the previous approaches we compare with.

**Existing System:**

There is already some work on reducing semantic drift. For example, Type Checking checks whether the type of extracted instances matches the target class, and Mutual Exclusion detects errors if extracted instances belong to mutually exclusive classes, such as “Animal” and “Artefact.”.

However, such constraints only tackle a small percentage of semantic drifts. Other methods keep the most reliable instances in each iteration to maintain high precision, where an instance’s “reliability” is determined either by some heuristic models (e.g., an instance is more reliable if it is extracted more frequently), or by combining evidences from multiple extractions. Not surprisingly, these methods sacrifice recall for increased precision.

**Proposed System:**

We present a novel approach to overcome semantic drift. We consider semantic drift to be triggered1 by certain patterns or instances that we call Drifting Points (DPs). The DPs themselves are not necessarily erroneous extractions, rather, they trigger semantic drift from the target class to some other classes such as the pattern P2 and the instance “chicken”. In that sense, erroneous extractions are the “symptoms” of semantic drift, while DPs are the “causes” of semantic drift. Identifying DPs enables us to cut off the propagation of semantic drift.

Compared with detecting each erroneous extraction directly, focusing on DPs makes the problem much easier, as DPs are easier to model for two reasons: First, the number of DPs is much smaller than that of erroneous extractions, as one DP may introduce many errors. Second, there are various kinds of erroneous extractions, which are hard to be captured by a single approach or model. In contrast, we identify two types of DPs that hold some strong features, which enable us to identify DPs and eventually identify erroneous extractions more effectively.