# Coupled Semi-Supervised Learning for Information Extraction

*Carlos Perez*

## 1 Introduction

The thesis of this paper is that it is possible to achieve high accuracy in semi-supervised learning by coupling the simultaneous training of many extractors. This paper advocates for semi-supervised learning as the cost associated with the training of the extractors for supervised learning is very high.

### Paper's contributions

This paper is the first to couple the semi supervised training of category and relation extractor and to train multiple wrapper inducers by using mutual exclusion and type checking relationships. It is also the first to propose the coupling of the training of freeform extractors and semi structured inducers.

## 2. Related work

There is a focus on bootstrapping methods, these, start with a small number of labeled “seed” examples and iteratively grow the set of labeled examples using high confidence labels from the current model. There is a problem called “semantic drift” and is the decline of the accuracy of bootstrapping methods after many iterations.

## 3. Coupled training

### 3.2 Coupling constraints

Coupled training constraints the iterative training to improve accuracy. There are three types of constraints considered in this paper

The first are mutual exclusion constraints that return a list of predicates that are mutually exclusive. These predicates cannot both be satisfied by the same input.

The second are Relation Argument Type Checking constraints that couple the learning of relation extractors with the learning of category extractors, using type checking.
In third place are Unstructured and Semi-structured Text Features constraints. These combine the results of classifiers that act on freeform contexts or semi structured contexts. Sirio can appear in freeform as “arg1 is the main star of”. Also this instance can appear in semi structured contexts such as HTML tags for a list item at a particular URL (A table with a list of start). The independent distribution of classifiers is assumed.

## 4 Algorithms

To address the problem of information extraction with high accuracy three algorithms are proposed CPL, CSEAL and MBL. These algorithms take as input a large text corpus, an initial ontology with predefined categories, relations, mutual-exclusion relationships and seed instance for all predicates.

### 4.1Coupled Patten Learner

CPL is a bootstrapping algorithm that leverages mutual-exclusion and type checking constraints to learn patterns that are accurate extractors of predicate instances.

#### 4.1.1 Extract candidates:

The first step in CPL is candidate extraction. Candidates can be of four types, category instances, category patterns, relation instances and relation patterns.

To extract category instances in the blank of a category pattern CPL looks for a noun phrase. For the sentene “There is a planet known as Nebulos” and pattern “planet known as arg1” Nebulos is extracted as a category instance candidate

To extract category patterns when a promoted category instance is found CPL extracts the preceding or following words according to some rules such as preceding words are are verbs followed by a sequence of adjectives (e.g.,“being attracted by arg1” or “asteroids attracted by arg1”)

To extract relation instances if a promoted relation pattern is found (e.g., “arg1 orbits around arg2”), a candidate relation instance is extracted if both placeholders are valid and they obey specifications for their categories.

To extract relation patterns if two arguments from a relation instance are found in a sentence the intervening sequence of words is extracted as a candidate relation pattern if it does not contain no more than five tokens has a content word and has an uncapitalized word. For example for category instances, Sirio of type star and Neptune of type planet, sentence, “Neptune will collide with Sirio”, it would extract the relation pattern “arg1 will collide with arg2”.

### 4.1.2 Filter candidates using coupling

Category instances are rejected unless the number of times they co-occur with a promoted pattern is at least three times more the time it co-occurs with patterns from mutually exclusive predicates. Pattern instances are filtered in a similar way using promoted category instances.

### 4.1.3 rank candidates

Category instances that have more co-occurrences with promoted patterns are ranked higher. Candidate patterns are ranked by using an estimate of their precision (p) higher p gives higher ranking. P is calculated as the number of times the pattern co-occurs with candidate categories over the number of times the pattern appears in the corpus.

### 4.1.4 Promote candidates

For each predicate CPL promotes at most 100 instances and 5 patterns according to the rankings. Category instances are only promoted if they co-occur with at least two promoted patterns. Pattern instances are promoted only if they co-occur with at least two category instances. Relation instances are only promoted if their arguments are candidates for the specified categories.

### 4.1.5 Large-Scale Implementation

CPL gathers statistics from a preprocessed text corpus which specifies how many times each noun phrase occurs with each category pattern and how many times each noun phrase occurs with each relation pattern. Preprocessing is done using MapReduce. In each iteration CPL gathers corpus statistics by scanning through the preprocessed data in two passes,one for extracting candidates and one for counting co-occurrences.

### 4.1.6 Uncoupled pattern learner

In the experimental part they use a variant of CPL called UPL which removes the coupling constraints from CPL .

## 4.2 Coupled SEAL

CSEAL that learns “wrappers” to extract instances from semi-structured documents and exploits mutual-exclusion and type checking constraints

CSEAL is based on SEAL. SEAL accepts input elements (seeds) of some target set S and automatically finds other probable elements of S. A category wrapper is extracted and CSEAL adds the constraints. The algorithm returns a list of candidate instances and the documents they were extracted from. CSEAL filters out documents that extract candidate instances from mutually exclusive predicates and only considers candidate relations if their arguments are candidate instances for their respective category types. It then ranks the candidate instances by the number of wrappers that extracted them and promotes the top 100 that were extracted by at least two wrappers.

## 4.3 Meta- Bootstrap- Learner

CPL and CSEAL are the subordinate algorithms of MBL. They do not promote instances on their own instead MBL promotes any instance that has been recommended by both techniques while obeying the mutual exclusion and type checking constraints specified in the ontology.

# 5 Experimental evaluation

### 5.1.1 Input Ontology

Input ontology contained categories and relations from two domains. Categories were initialized with 15 instances and 5 seed patterns. Relations were initialized with 15 instances and 5 negative instances and no seed patterns. Corpus was 200 million webpages and 514 million sentences.

## 5.3 Multiple Extraction Techniques

MBL has the highest precision of promoted instances out of all the algorithms considered. Coupling CPL and CSEAL with a multi-view coupling constraint yields more accurate learning than either method used alone.

# 6. Conclusion

They have presented methods that couple the semi supervised learning of category and relation instances and demonstrated that coupling forestalls the problem of semantic drift associated with bootstrap learning methods. The empirical evidence leads the authors to advocate for large-scale coupled training as a strategy to significantly improve accuracy in semi-supervised learning.