In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 115–118, Lisbon, Portugal, 2000.

Learning IE Rules for a Set of Related Concepts

J. Turmo and **H. Rodríguez**
TALP Research Center. Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3
E-08034 Barcelona - Spain

1 Introduction

The growing availability of on-line text has led to an increase in the use of automatic knowledge acquisition approaches from textual data. In fact, a number of Information Extraction (IE) systems has emerged in the past few years in relation to the MUC conferences[1]. The aim of an IE system consists in automatically extracting pieces of information from text, being this information relevant for a set of prescribed concepts (scenario). One of the main drawbacks of applying IE systems is the high cost involved in manually adapting them to new domains and text styles.

In recent years, a variety of Machine Learning (ML) techniques has been used to improve the portability of IE systems to new domains, as in SRV (Freitag, 1998), RAPIER (Califf and Mooney, 1997), LIEP (Huffman, 1996), CRYSTAL (Soderland et al., 1995) and WHISK (Soderland, 1999) . However, some drawbacks remain in the portability of these systems: a) existing systems generally depend on the supported text style and learn IE-rules either for structured texts, semi-structured texts or free text , b) IE systems are mostly single-concept learning systems, c) consequently, an extractor (e.g., a rule set) is learned for each concept within the scenario in an independent manner, d) the order of execution of the learners is set manually, and so are the scheduling and way of combination of the resulting extractors, and e) focusing on the training data, the size of available training corpora can be inadequate to accurately learn extractors for all the concepts within the scenario[2].

[1] http://www.muc.saic.com/

[2] This is so when dealing with some combinations of text style and domain.

This paper describes EVIUS, a multi-concept learning system for free text that follows a multi-strategy constructive learning approach (MCL) (Michalshi, 1993) and supports insufficient amounts of training corpora. EVIUS is a component of a multilingual IE system, M-TURBIO (Turmo et al., 1999).

2 EVIUS. Learning rule sets for a set of related concepts

The input of EVIUS is both a partially-parsed semantically-tagged[3] training corpus and a description of the desired target structure. This description is provided as a set of concepts $\mathcal{C}$ related to a set of asymmetric binary relations, $\mathcal{R}$.

In order to learn set $\mathcal{S}$ of IE rule sets for the whole $\mathcal{C}$, EVIUS uses an MCL approach integrating constructive learning, closed-loop learning and deductive restructuring (Ko, 1998).

In this multi-concept situation, the system determines which concepts to learn and, later, incrementally updates $\mathcal{S}$. This can be relatively straightforward when using knowledge about the target structure in a closed-loop learning approach. Starting with $\mathcal{C}$, EVIUS reduces set $\mathcal{U}$ of unlearned concepts iteratively by selecting subset $\mathcal{P} \subseteq \mathcal{U}$ formed by the *primitive concepts* in $\mathcal{U}$ and learning a rule set for each $c \in \mathcal{P}$ [4].

For instance, the single colour scenario[5] in fig-

[3] With EuroWordNet (http://www.hum.uva.nl/~ewn/) synsets. No attempt has been made to disambiguate such tags.

[4] No cyclic scenarios are allowed so that a topological sort of $\mathcal{C}$ is possible, which starts with a set of primitive concepts.

[5] Our testing domain is mycology. Texts consists of Spanish descriptions of specimens. There is a rich variety of colour descriptions including basic colours, intervals, changes, etc.

ure 1 is provided to learn from instances of the following three related concepts: *colour*, such as in instance "azul ligeramente claro" (slightly pale blue), *colour_interval*, as in "entre rosa y rojo sangre" (between pink and blood red), and *to_change*, as in "rojo vira a marrón" (red changes to brown).

Initially, $\mathcal{U} = \mathcal{C} =\{$*colour*, *colour_interval*, *to_change*$\}$. Then, EVIUS calculates $\mathcal{P} =\{$*colour*$\}$ and once a rule set has been learned for *colour*, the new $\mathcal{U} =\{$*colour_interval*, *to_change*$\}$ is studied identifying $\mathcal{P} = \mathcal{U}$.

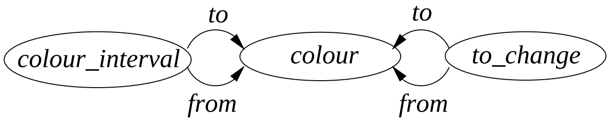

Figure 1: A single scenario for the colour domain

In order to learn a rule set for a concept, EVIUS uses the relational learning method explained in section 3, and defines the learning space by means of a dynamic predicate model. As a pre-process of the system, the training corpus is translated into predicates using the following initial predicate model: a) attributive meta-predicates: *pos_X*(A), *isa_X*(A), *has_hypernym_X*(A), *word_X*(A) and *lemma_X*(A), where X is instantiated with closed categories, b) relational meta-predicates: *distance_le_X*(A,B), stating that there are X terminal nodes, at most, between A and B, and c) relational predicates: *ancestor*(A,B), where B is the syntactic ancestor of A, and *brother*(A,B), where B is the right brother node of A sharing the syntactic ancestor.

Once a rule set for concept c is learned, new examples are added for further learning by means of a deductive restructuring approach: training examples are reduced to generate a more compact and useful knowledge of the learned concept. This is achieved by using the induced rule set and a syntactico-semantic transformational grammar. Further to all this, a new predicate *isa_c* is added to the model.

For instance, in figure 2[6], the Spanish sentence "su color rojo vira a marrón oscuro" (its red colour changes to dark brown) has two examples of *colour*, n3 and n6+n7, being these "rojo" (red) and "marrón"+"oscuro" (dark brown). No reduction is required by the former. However, the latter example is reduced to node n6'. As a consequence, two new attributes are added to the model: *isa_colour*(n3) and *isa_colour*(n6'). This new knowledge will be used to learn the concepts *to_change* and *colour_interval*.

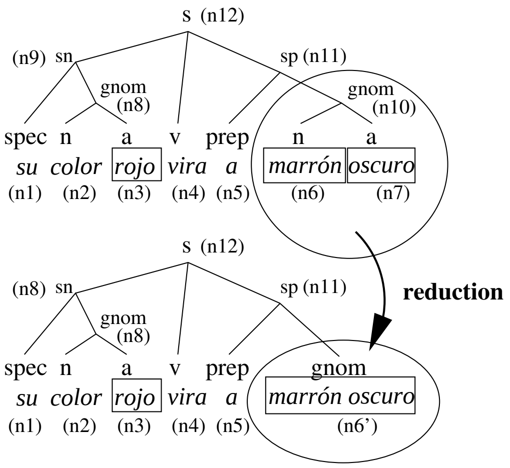

Figure 2: Restructuring training examples

3 Rule set learning

EVIUS uses FOIL (First-order Induction Learning) (Quinlan, 1990) to build an initial rule set $\mathcal{R}_0$ from a set of positive and negative examples. Positive examples $\mathcal{E}^+$ can be selected using a friendly environment either as:

- *text relations*: $c(A_1,A_2)$ where both A_1 and A_2 are terminal nodes that exactly delimit a text value for c. For instance, both text relations *colour*(n3,n3) or *colour*(n6,n7) in figure 2, or as:
- *ontology relations*: $c(A_1,A_2,\ldots,A_n)$ where all A_i are terminal nodes which are instances of already learned concepts related to c in the scenario. For instance, the ontology relation *to_change*(n3,n6')[7], in the same figure, means that the colour represented by instance n3 changes to that represented by n6'.

Negative examples $\mathcal{E}^-$ are automatically selected as explained in section 3.1.

[6]Which is presented here as a partially-parsed tree for simplicity.

[7]Note that, after the deductive restructuring step, both n3 and n6' are instances of the concept *colour*.

If any uncovered examples set, $\mathcal{E}_u^+$, remains after FOIL's performance, this is due to the lack of sufficient examples. Thus, the system tries to improve recall by growing set $\mathcal{E}^+$ with artificial examples (pseudo-examples), as explained in 3.2. A new execution of FOIL is done by using the new $\mathcal{E}^+$. The resulting rule set $\mathcal{R}'_0$ is combined with $\mathcal{R}_0$ in order to create $\mathcal{R}_1$ by appending the new rules from $\mathcal{R}'_0$ to $\mathcal{R}_0$. Consequently, the recall value of $\mathcal{R}_1$ is forced to be at least equal to that of $\mathcal{R}_0$, although the accuracy can decrease. A better method seems to be the merging of rules from $\mathcal{R}'_0$ and $\mathcal{R}_0$ by studying empirical subsumptions. This last combination allows to create more compact and accurate rule sets.

EVIUS uses an incremental learning approach to learn rule sets for each concept. This is done by iterating the process above while uncovered examples remain and the F_1 score increment ($\triangle F_1$) is greater than pre-defined constant α:

select $\mathcal{E}^+$ and *generate* $\mathcal{E}^-$
$\mathcal{R}_0 = \text{FOIL}(\mathcal{E}^+,\mathcal{E}^-)$
$\mathcal{E}_u^+ = \mathit{uncovered\_from}(\mathcal{R}_0)$
$\triangle F_1 = F_1(\mathcal{R}_0)$
while $\mathcal{E}_u^+ \neq \emptyset$ and $\triangle F_1 > \alpha$ *do*

$\mathcal{E}^+ = \mathcal{E}^+ \cup \mathit{pseudo\text{-}examples}(\mathcal{E}_u^+)$
$\mathcal{R}'_i = \text{FOIL}(\mathcal{E}^+,\mathcal{E}^-)$
$\mathcal{R}_{i+1} = \mathit{combine\_rules}(\mathcal{R}_i,\mathcal{R}'_i)$
$\mathcal{E}_u^+ = \mathit{uncovered\_from}(\mathcal{R}_{i+1})$
$\triangle F_1 = F_1(\mathcal{R}_{i+1}) - F_1(\mathcal{R}_i)$

endwhile
if $\triangle F_1 > \alpha$ *then return* $\mathcal{R}_{i+1}$
else return $\mathcal{R}_i$
endif

3.1 Generating relevant negative examples

Negative examples can be defined as any combination of terminal nodes out of $\mathcal{E}^+$. However, this approach produces an extremely large number of examples, out of which only a small subset is relevant to learn the concept. Related to this, (Freitag, 1998) uses words to learn only slot rules (learned from text-relation examples) , selecting as negative those non-positive word pairs that define a string as neither longer than the maximum length in positive examples, nor shorter than the minimum.

A more general approach is adopted to define the distance between possible examples in the learning space, applying a clustering method using positive examples as *medoids*[8]. The N nearest non-positive examples to each medoid can be selected as negative ones. Distance, in our case, must be defined as multidimensional due to the typology of occurring features. It is relatively easy to define distances between examples for *word_X* and *lemma_X* predicates, being 1 when X values are equal, and 0 otherwise. For *isa_X* predicates, the minimum of all possible conceptual distances (Agirre and Rigau, 1995) between X values in EWN has been used. Greater difficulty is encountered when defining a distance from a morpho-syntactic point of view (e.g., a pronoun seems to be closer to a noun than a verb). In (Turmo et al., 1999), the concept of δ*-set* has been presented as a syntactic relation generalization, and a distance measure has been based on this concept.

3.2 Creating pseudo-examples

A method has been used inspired by the generation of convex pseudo data (Breiman, 1998), in which a similar process to *gene-combination* in genetic algorithms is used.

For each positive example $c(A_1,\ldots,A_n)$[9] of concept c to be dealt with, an attribute vector is defined as

$$(\mathit{word\_X}_{B_1},\ldots,\mathit{word\_X}_{B_n},\mathit{lemma\_X}_{B_1},\ldots,\mathit{lemma\_X}_{B_n},\mathit{sem\_X}_{B_1},\ldots,\mathit{sem\_X}_{B_n},\mathit{context})$$

where $B_1,\ldots,B_n$ are the unrepeated terminal nodes from $A_1,\ldots,A_n$, *context* is the set of all predicates subsumed by the syntactico-semantic structure between the nearest positive example on the left and the nearest one on the right, and $\mathit{sem\_X}_{B_i}$ is the list of *isa_X* and *has_hypernym_X* predicates for B_i.

Then, for each example uncovered by the rule set learned by FOIL, a set of pseudo-examples is generated. A pseudo-example is built by combining both the uncovered example vector and a randomly selected covered one. This is done as follows: for each dimension, one of both possible values is randomly selected as value for the pseudo-example.

[8] A medoid is an actual data point representing a cluster.

[9] As defined in section 3.

T. Set*	$\mathcal{E}^+$	Recall	Prec.	F_1
15_0	105	56.86	100	0.725
25_0	206	62.74	98.45	0.766
35_0	270	73.53	97.40	0.838
45_0	328	75.49	98.72	0.856
55_0	398	75.49	98.72	0.856

Table 1: Results for the *colour* concept for different training set sizes (* subscript 0 means only one FOIL iteration)

4 Evaluation

EVIUS has been tested on the mycological domain. A set of 68 Spanish mycological documents (covering 9800 words corresponding to 1360 lemmas) has been used. 13 of them have been kept for testing and the others for training. The target ontology consisted of 14 concepts and 24 relations.

Several experiments have been carried out with different training sets. Results of the initial rule set for the *colour* concept[10] are presented in table 1.

Out of 34 in the 35_0 initial rule set, one of the most relevant learned rules is[11]:

$Colour(A, B)$:-$has\_hypernym\_00017586n(B)$, $has\_hypernym\_03464624n(A)$, $brother(A, B)$.

Table 2 shows the results of adding pseudo-examples to the 35_0[12] training set and using the algorithm in section 3. This was tested with $\alpha = 0.01$ (two iterations are enough, 35_1 and 35_2) and 5 pseudo-examples for each uncovered case. The algorithm returns the rule set produced in the first iteration due to the fact that $\triangle F_{1_T}$[13]> 0.01 between the first and the second iterations. Higher results can be generated when using lower values for α.

Although no direct comparison with other systems is possible due to the domain and language used, our results can be considered state-of-the-art regarding similar MUC competition tasks.

[10]This concept appears to be the most difficult to be learned.

[11]A chromatic colour (03464624n) that is the left syntactic brother of an attribute (00017586n) such as luminosity or another chromatic colour.

[12]This size has been selected to allow a better comparison with the results in table 1.

[13]F_{1_T} means the F_1 value for training sets

T. Set	$\mathcal{E}^+$	F_{1_T}	Recall	Prec.	F_1
35_1	415	0.981	76.47	97.50	0.857
35_2	465	0.987	79.41	97.50	0.875

Table 2: Results from adding pseudo-examples to the initial training set with 35 documents.

References

Eneko Agirre and German Rigau. 1995. A Proposal for Word Sense Disambiguation using Conceptual Distance. In *Proceedings of the International Conference RANLP*, Tzigov Chark, Bulgaria.

L. Breiman. 1998. Arcing Classifiers. *The Annals of Statistics*, 26(3):801–849.

M.E. Califf and R. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In *Workshop on Natural Language Learning*, pages 9–15. ACL.

D. Freitag. 1998. *Machine Learning for Information Extraction in Informal Domains.* Ph.D. thesis, Computer Science Department. Carnegie Mellon University.

S. Huffman. 1996. Learning information extraction patterns from examples. In S. Wermter, E. Riloff, and G. Sheller, editors, *Connectionist, statistical and symbolic approaches to learning for natural language processing.* Springer-Verlag.

H. Ko. 1998. Empirical assembly sequence planning: A multistrategy constructive learning approach. In I. Bratko R. S. Michalsky and M. Kubat, editors, *Machine Learning and Data Mining.* John Wiley & Sons LTD.

R.S. Michalshi. 1993. Towards a unified theory of learning: Multistrategy task-adaptive learning. In B.G. Buchanan and D. Wilkins, editors, *Readings in Knowledge Acquisition and Learning.* Morgan Kauffman.

J.R. Quinlan. 1990. Learning logical definitions from relations. *Machine Learning*, 5:239–266.

S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. 1995. Crystal: Inducing a conceptual dictionary. In *XIV International Joint Conference on Artificial Intelligence*, pages 1314–1321.

S. Soderland. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34:233–272.

J. Turmo, N. Català, and H. Rodríguez. 1999. An adaptable ie system to new domains. *Applied Intelligence*, 10(2/3):225–246.