Structuring Natural Language Data by Learning Rewriting Rules

Guillaume Cleuziou, Lionel Martin, and Christel Vrain

LIFO, Laboratoire d'Informatique Fondamentale d'Orléans
Rue Léonard de Vinci B.P. 6759
45067 Orléans cedex2 - France
{Guillaume.Cleuziou,Lionel.Martin,Christel.Vrain}@univ-orleans.fr

Abstract. The discovery of relationships between concepts is a crucial point in ontology learning (OL). In most cases, OL is achieved from a collection of domain-specific texts, describing the concepts of the domain and their relationships. A natural way to represent the description associated to a particular text is to use a structured term (or tree). We present a method for learning transformation rules, rewriting natural language texts into trees, where the input examples are couples (*text, tree*). The learning process produces an ordered set of rules such that, applying these rules to a *text* gives the corresponding *tree*.

1 Introduction

The work presented in this paper has been motivated by a French project (ACI Biotim http://www-rocq.inria.fr/imedia/biotim/) in the field of Biodiversity. The task we address aims at semi-automatically building an ontology of the domain from corpora describing flora.

The term *ontology* has various definitions in various domains. From a practical point of view, an ontology can be defined as a quadruple $O = (C, R, A, Top)$ where C is a set of concepts, R is a set of relations, A is a set of axioms and Top is the highest-level concept [SB03]. The set R contains relations between concepts, as for example, the binary relation *partof* relating the concepts *hand* and *human*. Usually we distinguish taxonomic and non-taxonomic relations: taxonomic relations are used to organize information with generalization/specialization (or hyponymy) relationships in a "ISA hierarchy"; non-taxonomic relations are any other relations such as synonymy, meronymy, antonymy, attribute-of, possession, causality, ...

Ontology learning refers to extracting one of these elements from input data. This task has been addressed in several research areas. Ontology learning systems extract their knowledge from different types of sources, such as structured data (databases, existing ontologies, ...) or semi-structured data (dictionaries, XML documents, ...). One of the problems is to learn from unstructured data

S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad (Eds.): ILP 2006, LNAI 4455, pp. 125–138, 2007.

(domain-specific natural language texts). A quite natural formalism for structuring texts is first-order logics (usually logic programs), thus allowing the use of Inductive Logic Programming for different tasks, as for instance Text Categorization, Information Extraction or Parser Acquisition [Coh95, JSR99, Moo96]. This usually leads to a two-step process: a syntactic analysis of the texts, followed by the learning task. Nevertheless, in our application, the corpora is specific (long descriptions of flore without verbs) making difficult the use of classical syntactic parsers. For instance, the following example is the beginning of the description of the plant called "Pulchranthus variegatus":

> *" Subshrubs or shrubs, 0.5-2 m tall. Stems terete with red, exfoliating bark. Leaves: petioles 3-13 mm long; blades elliptic-lanceolate, 13-26 × 5-9 cm, glabrous, the apex acuminate-cuspidate. Inflorescences terminal, racemes or panicles, 4-15 cm long, green, the flowers 2-many per node; peduncle 10-15 mm long; bracts small, narrowly triangular, 2.5-3 x 0.5 mm; pedicels lacking to short, 1.5 mm long; bracteoles 1.5-2 mm long. ..."*

This text describes different concepts (stem, bark, leaf, ...) and various relations: part-of relations (bark is a part of a stem, flower is a part of inflorescences, ...) and attribute-value relations (stem is terete, bark is red, petiole is 3-13 mm long...).

All this information can be represented into a tree (term), the leaves (constants) are elements of the text. For example, the term

partOf(desc(stem, terete), desc(bark, [red, exfoliating]))

could be a representation of information associated with the sentence *"Stems terete with red, exfoliating bark"*. The detailed formal language used in our work is presented in Section 3.

Given a set of sentences and their corresponding terms (manually built), our goal is to produce a set of rules able to rewrite a sentence into a term. The corpora shows that in many cases, some simple regular structures can be automatically discovered, these structures are based on the punctuation and the syntactical categories of words. For example, when a noun is immediately followed by an adjective, then the adjective describes the noun; when two descriptions are separated by "," or "*, with*", then the second description is about a concept which is a part of the concept of the first description. This short example also shows that a preprocessing step is required: the initial text is transformed into a list of elements (words, punctuation), each element is tagged, using a part-of-speech (POS) tagger; this preprocessing is done in most existing ontology learners.

Some works have already adressed this task: [MPS02] proposes a survey of methods relying either on statistics or predefined patterns, [SM06] is based on cooccurrences with verb phrases, [Yam01] uses a n-grams representation and [Ait02] uses ILP techniques to characterize specific relations. [Bri93] proposes a transformation-based approach for parsing text into binary trees.