PACTOLE: A methodology and a system for semi-automatically enriching an ontology from a collection of texts

Rokia Bendaoud, Yannick Toussaint, Amedeo Napoli

PACTOLE: A methodology and a system for semi-automatically enriching an ontology from a collection of texts

Rokia Bendaoud, Yannick Toussaint, and Amedeo Napoli

UMR 7503 LORIA, BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE
{bendaoud,yannick,napoli}@loria.fr

Abstract. PACTOLE stands for "Property And Class characterization from Text for OntoLogy Enrichment" and is a semi-automatic methodology for enriching an initial ontology from a collection of texts in a given domain. PACTOLE is also the name of the associated system relying on Formal Concept Analysis (FCA). In this way, PACTOLE is able to derive a concept lattice from a formal context, consisting of a binary table describing a set of individuals with their properties. Given a domain ontology and a set of objects with their properties (extracted from a collection of texts), the PACTOLE system builds two concept lattices: the first corresponding to the restriction of the ontology schema to the considered objects and the second to the extracted pairs (object, property). As they are based on the same set of individuals, the two ontologies are merged using context apposition. The resulting final concept lattice is analyzed and a number of knowledge units can be extracted and furthermore used for enriching the initial ontology. Finally, the final concept lattice is mapped within the $\mathcal{FLE}$ KR formalism. The paper introduces and explains in details the PACTOLE methodology with the help of an example in the domain of astronomy.

1 Introduction

1.1 Motivation and context

Ontologies are the backbone of Semantic Web. They help software and human agents to communicate by providing shared and common domain knowledge, and by supporting various tasks, e.g. problem-solving and information retrieval [11]. An ontology is usually based on a concept hierarchy and a set of relations between the concepts. In turn, a concept hierarchy structures domain knowledge into a set of hierarchically organized classes, making easier information search and reuse. However, the design and the enrichment of an ontology are hard and time-expensive tasks. Indeed, the knowledge acquisition bottleneck is one major factor slowing down ontology-driven applications [3]. This point is illustrated hereafter by an example taken from the domain of astronomy and used in the whole paper (this research work is carried out in the context of a project done in collaboration with researchers in astronomy). In this application domain, the

design of a concept hierarchy and the identification/classification of celestial bodies, i.e. assigning a class to a given celestial body, are very difficult tasks, because of the growing number of discovered celestial bodies and the need of new classes to be defined. Traditionally, the classification task is performed "manually", according to the object properties appearing in the astronomy documents. The task consists in firstly reading scientific articles holding on the celestial object under study and secondly finding a possible class for that object. At present, more than three millions of celestial objects are classified in this way and made available in the SIMBAD database[1]. The SIMBAD database is one of the most important databases in astronomy memorizing the properties of celestial objects. But the SIMBAD database remains a database and has not the architecture of an ontology: no definition, no explicit representation of relations, no classification procedures built-in, and a considerable work has to be done for classifying the billion of remaining celestial objects. The task is tedious for human experts, who are not always confident with their own classification, mainly because classes lack precise and unambiguous definitions. Thus, the design of an ontology for guiding the classification of celestial bodies would be of great help for astronomy practitioners.

In this way, this paper presents a methodology and a system for designing an ontology from a collection of astronomical texts. One originality is that the resulting ontology is completed with the help of domain resources, e.g. domain ontology, database, or thesaurus. Accordingly, and this is the case in this paper, the methodology can be used for enriching the knowledge included in an existing resource, here a domain ontology based on the SIMBAD database. This approach can be used for partly solving the knowledge acquisition bottleneck. Moreover, it can be noticed that the methodology is not dependent on the domain and other experimentations have been carried out in the domain of biology. More precisely, the PACTOLE methodology –PACTOLE stands for "Property And Class Characterization from Text to OntoLogy Enrichment"– takes as input a collection of texts in astronomy and a domain resource, i.e. an ontology based on the SIMBAD database, and gives as output a set of new concepts and instances to be inserted in the initial ontology. The enrichment process is based on Formal Concept Analysis (FCA) [7]. In addition, for being inserted in the ontology, all knowledge units are represented within the Description Logics (DL) language $\mathcal{FLE}$ where the following constructors are available: conjunction ($\sqcap$), universal quantification ($\forall$), and existential quantification ($\exists$). The description logics $\mathcal{FLE}$ is used for representing concepts and relations in the ontology and has a sufficient power of representation for that task.

Actually, the PACTOLE system implements the PACTOLE methodology and builds two concept hierarchies using FCA: one concept hierarchy derives from the collection of texts and one concept hierarchy derives from the SIMBAD database (mentioned here before as the ontology based on the SIMBAD database). After that, the two concept hierarchies are merged by the operation of context apposition as introduced and discussed in [7].

[1] `http://simbad.u-strasbg.fr/simbad/sim-fid`

Applying in this way the FCA process for the enrichment of an ontology is an original design operation that brings forward two main benefits. Firstly, a FCA-based concept hierarchy provides a formal basis and specification for the resulting ontology. Moreover, many efficient FCA-based operations are designed for extending, maintaining, and managing a concept hierarchy, such as performing an incremental update of the hierarchy by adding either an object or an attribute (property), or assembling a concept lattice from parts. Secondly, as the concept hierarchy changes (because texts are changing for example), the ontology evolves in a correct and consistent way. The transformation of the concept lattice into a DL knowledge base (KB) allows then to query the KB with the help of a DL reasoner and to ask complex expert questions.

1.2 An introductory example

Let us consider the problem of detecting why two celestial objects are in the same class. To answer the question, the set of properties shared by both objects has to be characterized. The extraction of such set of common properties relies on a search in astronomical texts of elements that can be considered as properties for identifying the class of an object. For example, in a sentence such as "*We report the discovery of strong flaring of the object* `HR2517`", it is asserted that the object `HR2517` can *flare*, i.e. showing an eruption of plasma at the surface of the object. The fact of flaring means for a celestial object, here `HR2517`, that the object is a particular type of star. In another sentence such as "`NGC 1818` *contains almost as many Be stars as the slightly younger* `SMC` *cluster* `NGC 330`", it is asserted that the object `NGC 1818` *contains* something. The fact of containing means that this celestial object is not a star.

In these sentences, the property of an object is given by a verb. A similar approach has been used in [6] and is based on Harris hypothesis [10], stating that terms in sentences are similar if they share similar linguistic contexts, here the similarity of verb-argument dependencies. In this way, individuals and their properties are extracted from a collection of texts using Natural language processing (NLP) tools. Then, the FCA process is used for building a concept hierarchy from a formal context, composed of a set of individuals, e.g. `SMC`, `T`, `Tauri`, a set of properties, e.g. `contains`, `flaring`, and a binary relation defined on the Cartesian product of both sets stating that an object has or has not a given property.

Given a concept hierarchy and the derived ontology represented in the $\mathcal{FLE}$ DL, complex expert questions can be answered. The questions are first given in natural language and then represented as DL queries. Such expert questions can be read as the following: *do the celestial objects* `3C 273` *and* `SMC` *belong to the same class?* or *What is the class of the celestial object* `V773 Tau`*?*.

1.3 Organization of the paper

The following sections of this paper are organized as follows. The next section introduces the definitions of ontology enrichment and the basics of FCA. In

the Section 3, the PACTOLE methodology is presented and the operations of knowledge extraction from texts, concept hierarchy design and representation (in $\mathcal{FLE}$), and ontology enrichment, are explained and illustrated. In Section 4, an evaluation of each step of the PACTOLE methodology system is given followed by a discussion and a synthesis of the present research work. Section 5 briefly presents related works on ontology design and enrichment. Finally, the Section 6 concludes the paper and shows future works.

2 Ontologies enrichment and Formal Concept Analysis

In this section, the background definitions for the PACTOLE methodology are given. According to the general and commonly admitted statement in [9], an ontology is an explicit specification of a domain conceptualization. Moreover, an ontology is usually developed for the purposes of domain knowledge sharing and reuse. Following this way, the objective of the PACTOLE methodology is to enrich an existing domain ontology from a collection of texts, to solve a particular problem, e.g. expert question answering.

2.1 The enrichment of an ontology

The following definition of ontology enrichment is based on the work of Faatz and Steinmetz [5]. This enrichment operation is based on a so-called "set of formulas" for each concept of the initial ontology, including new concepts, new properties, and new instances.

Definition 1 (Ontology Enrichment). *Let Texts be a collection of written texts and Exp(Texts) a set of expressions that have been extracted from Texts by NLP tools. Expressions may be nouns or pairs (subject, verb). An algorithm for ontology enrichment from text denoted hereafter by AOET takes as input an ontology Ω and a set Exp(Texts), and returns as output an enriched ontology $\Omega \cup P$, where P is a set of formulas represented within the same representation formalism as Ω and obtained as follows. For each element $e \in$ Exp(Texts), AOET returns a formula f(e) that can be either an individual, a concept, or a role, involving e, and depending on the status of e in Exp(Texts), as explained in the following.*

2.2 Formal Concept Analysis

Formal concept analysis (FCA) [7] is a mathematical formalism allowing to derive a concept lattice (to be defined later) from a formal context $\mathbb{K}$ constituted of a set of objects G, a set of attributes M, and a binary relation I defined on the Cartesian product $G \times M$ (in the binary table representing $G \times M$, the rows correspond to objects and the columns to attributes or properties). FCA can be used for a number of purposes among which knowledge formalization and acquisition, ontology design, and data mining. The concept lattice is composed

of *formal concepts*, or simply *concepts*, organized into a hierarchy by a partial ordering (a subsumption relation allowing to compare concepts). Intuitively, a concept is a pair (A, B) where $A \subseteq G$, $B \subseteq M$, and A is the maximal set of objects sharing the whole set of attributes in B and vice-versa. The concepts in a concept lattice are computed on the basis of a Galois connection defined by two derivation operators denoted by $'$:

$$' : G \rightarrow M; A' = \{m \in M; \forall g \in A : (g, m) \in I\}$$
$$' : M \rightarrow G; B' = \{g \in G; \forall m \in B : (g, m) \in I\}$$

Formally, a concept (A, B) verifies $A' = B$ and $B' = A$. The set A is called the *extent* and the set B the *intent* of the concept (A, B). The subsumption (or subconcept–superconcept) relation between concepts is defined as follows: $(A_1, B_1) \sqsubseteq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2$ (*or* $B_2 \subseteq B_1$). Relying on this subsumption relation $\sqsubseteq$, the set of all concepts extracted from a context $\mathbb{K} = (G, M, I)$ is organized within a complete lattice, that means that for any set of concepts there is a smallest superconcept and a largest subconcept, called the *concept lattice* of $\mathbb{K}$ and denoted by $\underline{\mathfrak{B}}(G, M, I)$.

3 The PACTOLE Methodology

PACTOLE is a methodology for enriching in a semi-automatic way an initial ontology based on a domain resources (thesaurus, database,...) with knowledge extracted from texts. PACTOLE is inspired from two methodologies, namely "Methontology" [8] and "SENSUS" [14]. From "Methontology", PACTOLE borrows the idea of keeping an expert in the loop to validate operations such as building from a set of terms extracted from resources defining a set of DL concepts. From "SENSUS", PACTOLE borrows the idea of being based on an existing ontology and enriching this initial ontology with resources such as texts. The PACTOLE process is based on five steps presented in Figure 1, each step in PACTOLE involves the experts validation.

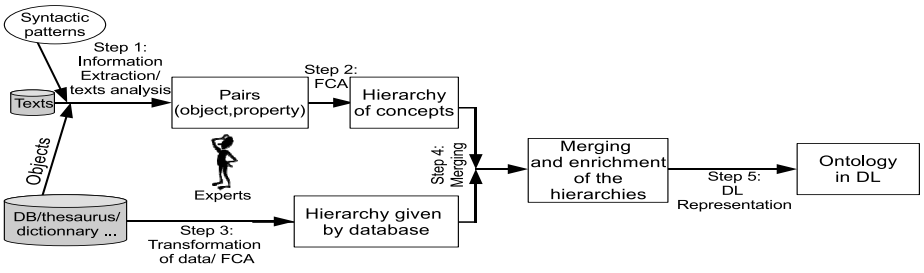

Fig. 1. PACTOLE Methodology

The first step involves NLP processing for extracting from texts objects of the domain and their properties. The expressions that are considered are

verb/subject, verb/object, verb/complement, and verb/prepositional phrase dependencies. They are good syntactic hints for assigning a property to an object. Each of these hints provides a pair (object, property). In the second step, FCA is used for building a concept lattice from the pairs (object, property). A concept in the hierarchy is composed of a maximal set of individuals sharing a maximal set of attributes (or properties) and vice-versa. The third step converts the existing knowledge resources into a lattice structure using FCA. During the fourth step, the two lattices are merged. The idea here is that the concept hierarchy from the initial knowledge resources can be partially enriched by the concept lattice resulting extracted from texts. During step five, the final (merged) lattice is represented with the $\mathcal{FLE}$ DL formalism. The following subsections give details on each step.

3.1 Text Analysis

This step aims at extracting from the texts a list of pairs (object, property). A preliminary task identifies celestial objects in the texts. Then, texts are parsed to extract syntactic dependencies, and some syntactic dependencies involving celestial objects are selected and translated into pairs of the form (celestial_object, property).

Detection of celestial objects. There is no normalization process for naming a celestial object in astronomy. Thus, identifying the names of the objects in the texts requires two complementary strategies which are suggested by the SIMBAD database: some names are already known (such as "`Orion`") and the string can be used to locate them in the texts. Some other names such as "`NGC 6994`" are described by a pattern "`NGC NNNN`" where NNNN is a number.

The system has extracted 1382 celestial objects from the collection of texts, this number representing 90% of the whole set of objects in the texts (as evaluated by the experts). Three new objects were identified: they were not in the SIMBAD database: `HH 24MMS, S140 IRS3, M33 X-9`. However, a few detected objects were not celestial objects. Three main failures in object identification have been pointed out:

- Underspecified patterns: some objects having the same pattern as celestial objects are not celestial objects: The `IRA X` pattern in SIMBAD covers `IRAS 16293` which is a celestial object but also `IRAM 30` which is a telescope,
- Abbreviations in texts: some authors use short ways to name objects in the texts, e.g. `S 180` instead of `Sand 180` as registered in SIMBAD,
- Typing errors in SIMBAD: some errors were made while typing the name of objects in SIMBAD, e.g. `Name Lupus 2` instead of `Lupus 2`.

Extraction of properties. The properties are extracted by parsing the texts with the shallow "Stanford Parser"[2] [4]. The Stanford Parser parses texts and

[2] http://nlp.stanford.edu/software/lex-parser.shtml

extracts syntactic dependencies between a verb and its subjects, objects, complements, and preposition phrases. For example: *"NGC 1818 contains almost as many Be stars as the slightly younger SMC cluster NGC 330"*. The list of dependencies is the following:

– subject(contains-2,NGC 1818-1), direct_object(contains-2,Be stars-6)

Only verb dependencies are kept to build the pairs (celestial_object, property). The pair (contains, NGC 1818) is derived from the dependency subject(contains-2, NGC 1818-1), meaning that `NGC 1818` is able to `contain`. The pair (Be star, contained) is derived from dependency direct_object(contains-2, Be stars-6), meaning that Be stars can be `contained`.

Among the set of pairs (object, property), some are pure linguistic artefacts. They are not relevant to astronomy and should be filtered before the classification process. Firstly, properties which occurs only once are considered as noise and deleted. Secondly, the system deals with synonymy (`consists`, `contains` and `includes`...) for reducing dispersion. These properties are grouped and considered as the same property. Finally, for each remaining pair, an astronomer decides whether it is meaningful to keep the pair for the classification process. For example, properties such as `performing` or `oscillating` have been considered of low interest, while some others pairs such as `rotating` were considered as interesting.

This step allows the system to discover some properties which were previously unknown, in the sense that no correlation was known between celestial types of objects and properties. For example, the objects "`59 Aurigae, V1208 Aql`" can `pulse`, the object "`MM Herculis`" can `eclipse` or the objects "`AB Dor, OJ 287`" can `flare`.

3.2 Classifying celestial objects from the texts using FCA

The set of pairs extracted from the text are then transformed under the form of a binary table objects $\times$ properties leading to a formal context $\mathbb{K}_1=(G, M_1, I_1)$ to which FCA method will be applied. Here G is a set of the celestial objects identified in the texts, M_1 is the set of properties extracted from texts and modified as described above, and I_1 is the relaion and $I_1(g, m_1)$ is a statement, that g has the property m_1. An example of such a lattice is given in Figure 2.

3.3 Classifying celestial objects from Simbad database using FCA

The hierarchical structure defined in SIMBAD is encoded into a concept lattice so that both hierarchical structures – from SIMBAD and from texts – are expressed in the same formalism namely a concept lattice. The context related to SIMBAD is $\mathbb{K}_2=(G, M_2, I_2)$ where G is a set of celestial objects identified in the texts, M_2 is the set of SIMBAD classes, and $I_2(g, m_2)$ is the relation stating that g has or has not the class m_2. An example of concept lattice extracted from SIMBAD is given on Figure 3.

	observed	expanding	flaring	emits	includes
3C 273	X			X	X
TWA	X	X			
SMC	X			X	
T Tauri	X		X	X	
V773 Tau	X		X	X	

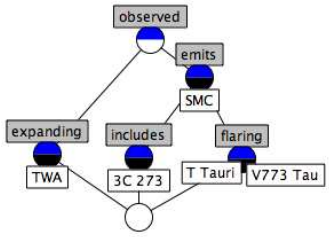

Fig. 2. The context $\mathbb{K}_1 = (G, M_1, I_1)$ and the lattice of this context

	Quasar	Association of Stars	Galaxy	Star	T Tau type Star
3C 273	X		X		
TWA		X			
SMC			X		
T Tauri				X	X
V773 Tau				X	X

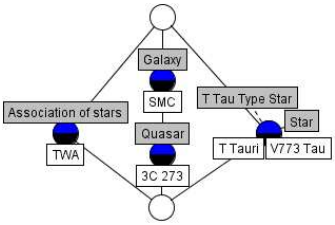

Fig. 3. The context $\mathbb{K}_2$=(G, M_2, I_2) and the lattice of this context

3.4 Merging the two lattices

The PACTOLE system proposes to enrich the lattice resulting from SIMBAD with the concept lattice of celestial objects built from the texts. Merging these two concept lattices relies on the apposition operation as defined in [7]:

Definition 2. *Let* $\mathbb{K}_1$=(G_1, M_1, I_1), *and* $\mathbb{K}_2$=(G_2, M_2, I_2) *be formal contexts. If* $G = G_1 = G_2$ *and* $M_1 \cup M_2 = \emptyset$ *then:* $\mathbb{K} := \mathbb{K}_1|\mathbb{K}_2 := (G, M_1 \cup M_2, I_1 \cup I_2)$ *is the apposition of the two contexts* $\mathbb{K}_1$ *and* $\mathbb{K}_2$.

The two contexts are respectively $\mathbb{K}_1$=(G, M_1, I_1) (presented in Figure 2) and $\mathbb{K}_2$=(G, M_2, I_2) (presented in the Figure 3). The apposition context $\mathbb{K} = (G, M, I)$ is presented in the Table 1 where G is the same set of objects for $\mathbb{K}_1$ and $\mathbb{K}_2$, $M := M_1 \cup M_2$ where M_1 is the set of properties extracted from the texts and M_2 is a set of the classes of SIMBAD, and $I := I_1 \cup I_2$. The resulting concept lattice is presented in Figure 4.

3.5 Representing the concepts with $\mathcal{FLE}$

The last step in PACTOLE is aimed at transforming the final lattice into an ontology represented in $\mathcal{FLE}$.

This transformation called α is based on a set of elementary transformations defined as follows: $\alpha : \mathbb{K} = (G, M, I) \rightarrow$ TBox $\sqcup$ ABox, where: $\mathbb{K}$ is a formal context, TBox and ABox being the bases of the ontology. The elementary transformations are the following:

Table 1. The context $\mathbb{K} = (G, M, I)$

	Quasar	Association of Stars	Galaxy	Star	T Tau type Star	observed	expanding	flaring	emits	includes
3C 273	X		X			X			X	X
TWA		X				X	X			
SMC			X			X			X	
T Tauri				X	X	X		X	X	
V773 Tau				X	X	X		X	X	

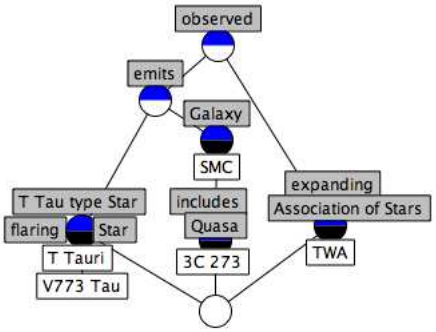

Fig. 4. Lattice of the context $\mathbb{K} = (G, M, I)$

1. A formal attribute $m_2 \in M_2$ is transformed in the TBox as an atomic concept c $\equiv \alpha(m_2) \equiv m_2$. A class in SIMBAD is represented as a concept, e.g. α(`quasar`)=`quasar`,
2. A formal attribute $m_1 \in M_1$ is transformed in the TBox as a defined concept c $\equiv \alpha(m_1) \equiv \exists m_1.\top$. Formal attributes are used as roles for defined concepts, e.g. α(`observed`) $\equiv \exists$`observed.`$\top$,
3. A formal concept $c = (X, Y) \in \mathcal{C}$ is transformed in the TBox as defined concept $\alpha(c)$, i.e. $\alpha(c) \equiv \sqcap_{m \in Y} \alpha(m)$ where $\alpha(m)$ are either atomic or defined concepts, e.g. α(`C`$_4$) $\equiv$ `Star` $\sqcap$ `T_Tau_type_Star` $\sqcap \exists$`observed.`$\top \sqcap \exists$`emits.`$\top \sqcap \exists$`flaring.`$\top$,
4. A subsumption relation between formal concepts C and D is transformed in the TBox as a general concept inclusion $\alpha(C) \sqsubseteq \alpha(\bar{D})$, e.g. α(`C`$_4$) $\sqsubseteq$ α(`C`$_1$),
5. A formal object $g \in G$ is transformed in the ABox as an instance $\alpha(g)$, e.g. α(`T Tauri`) = `T_Tauri` is an instance.

The definition of each concept of the final lattice in Figure 4 is presented in Table 2. The resulting ontology shown in Figure 5 can be used for two kinds of tasks:

1. Instantiation of concepts. Let o_1 be a celestial object having the properties {`a,b`} and belonging to classes {`C`$_1$`,C`$_2$} in SIMBAD. A first task is instantiation, i.e. finding the class of an object such as o_1. The class of o_1 is a most general class X in the final ontology such that X $\sqsubseteq \exists$a.$\top \sqcap \exists$b.$\top \sqcap$C$_1 \sqcap$

Table 2. Definition of each concept of the final lattice

N° in the lattice	Definition
C_0	$\exists$observed.$\top$
C_1	$\exists$observed.$\top \sqcap \exists$emits.$\top$
C_2	Association_of_Stars $\sqcap \exists$observed.$\top \sqcap \exists$expanding.$\top$
C_3	Galaxy $\sqcap \exists$observed.$\top \sqcap \exists$emits.$\top$
C_4	Star $\sqcap$ T_Tau_type_Star $\sqcap \exists$observed.$\top \sqcap \exists$emits.$\top \sqcap \exists$flaring.$\top$
C_5	The bottom : $\bot$
C_6	Galaxy $\sqcap$ Quasar $\sqcap \exists$observed.$\top \sqcap \exists$emits.$\top \sqcap \exists$includes.$\top$

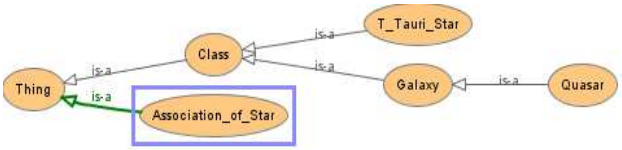

Fig. 5. Final ontology

C_2. When there exists more than one candidate class for being the class of an object o_1 say D_1 and D_2, the conjonction $D_1 \sqcap D_2$ becomes the class of o_1. For example, let us consider the question *"What is the class of the object V773 Tau, having the properties* {observed,flaring,emits} and belonging to the classes {Star,T Tau Star} in SIMBAD? The answer is the most general class $X \sqsubseteq \exists$observed.$\top \sqcap \exists$flaring.$\top \sqcap \exists$emits.$\top \sqcap$Star$\sqcap$T_Tau_Star, here the concept C_4 in the ontology.

2. Comparison of celestial objects. Let us consider two objects o_1 and o_2. A second task consists in comparing o_1 and o_2 and determining whether o_1 and o_2 are in the same class. One way for checking that is to find the class of o_1, then the class of o_2, and then to test whether the two classes are identical. For example, let us consider the two objects named 3C 273 and SMC. The object 3C 273 is an instance of the class C_6 and the object SMC is an instance of the class C_3. As $C_6 \sqcap C_3 = C_6$, the objects 3C 273 and SMC are not in the same class.

4 Evaluation

In this section, the PACTOLE methodology is evaluated, mainly by comparing the concept hierarchy associated to the resulting ontology and the initial existing hierarchy, here the SIMBAD database. The PACTOLE system has been applied on 11591 abstracts from the A&A "Astronomy and Astrophysics" journal for the years 1994 to 2002.

4.1 Evaluation of the process

The Stanford Parser analyzes 68.5% of the sentences in the texts, where the maximum size of the parsed sentences is between 31 and 36 words. The system extracts three different sets of syntactic dependencies between verb and arguments, namely SO, SOC, and SOCP (detailed in Table 3) where:

- SO: subject(object,verb) + object(object,verb),
- SOC: SO + complement(object,verb),
- SOCP: SOC + preposition_X(object,verb), where X can be (in, of, ...).

Table 3. The results of the parser

	SO				SOC				SOCP			
	Pairs	Obj.	Prop.	Conc.	Pairs	Obj.	Prop.	Conc.	Pairs	Obj.	Prop.	Conc.
11591 abstracts	384	209	14	30	401	211	14	30	1709	470	23	70

A concept lattice with 94 concepts has been built from the SIMBAD database, where 470 objects and 92 properties have been considered in the formal context.

The lattice resulting from apposition was presented to the astronomers. Actually, new concepts have been discovered such as the concept ({`Orion, TWA`}, {`Association_of_stars, expanding, observed`}). This concept represents the `Association_of_stars` than can `expand`. The concept is considered as interesting by domain experts, and labelled as the `Association_of_Young_Stars`.

4.2 Evaluation of hierarchy correspondence

The correspondence between the concept hierarchy extracted from the collection of texts and the concept hierarchy extracted from SIMBAD database has to be checked. Here the objective is to check whether the PACTOLE system has defined each class of the concept hierarchy resulting from SIMBAD (validation classes) as a class with properties extracted from the collection of texts (experimentation classes). This correspondence relies on similarity between sets of instances. In order to do so, the measures of precision and recall have been used. The precision and the recall are calculated for each experimentation classes with respect to one of the closest class in verification class using the Euclidean distance. The global precision (Precision_F) and the global recall (Recall_F) are the average of all precisions (respectively of all recalls).

Calculate the global precision and recall. The precision is the number of common instances between C_{E_i} (experimentation class i) and C_{V_j} (validation class j) divided by the number of instances in C_{E_i}. The recall is the number of common instances between C_{E_i} and C_{V_j} divided by the number of instances in C_{V_j}. N is the number of classes in C_E.

$$Precision_i = \frac{C_{E_i} \cap C_{V_j}}{C_{E_i}}, \quad Recall_i = \frac{C_{E_i} \cap C_{V_j}}{C_{V_j}}$$

$$Precision\_F = \frac{\sum_{i=1..N}(Precision_i)}{N}, \quad Recall\_F = \frac{\sum_{i=1..N}(Recall_i)}{N}$$

Detection of the closest class. For each class has been searched for one of the closest class in the classes of SIMBAD using the Euclidian distance, if we find two closest classes, one of them is taken. Let G be the set of objects, E the set of experimentation classes, and V a set of validation classes. For each class $C_{E_i} \in E$, and for each class $C_{V_j} \in V$, vector V_{E_i} and V_{V_j} are defined as:

$\forall g \in G$: if g is an instance of C_{E_i} then $V_{E_i}[g] = 1$ else $V_{E_i}[g_]] = 0$

$\forall g \in G$: if g is an instance of C_{V_i} then $V_{V_i}[g] = 1$ else $V_{V_i}[g] = 0$,

then:

$$Distance(V_{E_i}, V_{V_j}) = (\sum_{k=0}^{N}(V_{E_i}[g] - V_{V_j}[g])^2)^{1/2}$$

C_{V_j} is one of the closest class of C_{E_i} iff $\forall V_{V_p} \in V - \{V_{V_j}\}$ Distance $(V_{E_i}, V_{V_p}) \geqslant$ Distance (V_{E_i}, V_{V_j}).

For example, let G be the set of objects $G = \{$`3C 273, TWA, SMC, T Tauri, V773 Tau`$\}$ (see the Figures 2 and 3). One of the closest class for C_{E_1} with instances $\{$`3C 273, SMC, T Tauri, V773 Tau`$\}$ (Figure 2) is class C_{V_1} in SIMBAD with instances $\{$`3C 273, SMC`$\}$ (Figure 3). The distance between the vector associated to C_{E_1} that is V_{E_1} = [1,0,1,1,1] and the vector associated to C_{V_1} that is V_{V_1} = [1,0,1,0,0] is the minimal distance.

$Distance(V_{E_1}, V_{V_1}) = \sqrt{2}$, $Precision_1 = \frac{C_{E_1} \cap C_{V_1}}{C_{E_1}}$=0.5, $Recall_1 = \frac{C_{E_1} \cap C_{V_1}}{C_{V_1}}$=1.

Table 4. Resulting measures of precision and recall for differents set of dependencies

	SO		SOC		SOCP	
	Final Precision	Final Recall	Final Precision	Final Recall	Final Precision	Final Recall
FCA	58.33%	05.03%	58.91%	05.94%	74.71%	30.22%

4.3 Discussion

The PACTOLE system allows to extract new knowledge units in the astronomy domain and to enrich an ontology associated to the SIMBAD database. These knowledge units can be divided in three kinds. The first kind is related to the identification of new celestial objects (see the subsection 3.1). The second kind is related to the discovery of new correlations between celestial objects and their properties (see the subsection 3.1). The third kind is related to the proposition of new classes in SIMBAD (see the subsection 4.1). The experiment in astronomy shows also that using all syntactic dependencies (SOCP) leads to better results.

The SOCP set allows the extraction of more pairs, more properties and more classes (see Table 3). This set also offers a better precision and a better recall

(see Table 4). The score of precision is high (74.71%) meaning that objects are classified in adequate classes. The score of recall is low for several reasons. The first reason is that the number of properties associated with objects is not sufficient. Sometimes, the system extracts only one or two properties for an object and this is too small for classification. The second reason is that verbs are not the sole properties for defining a class, considering for example adjectives, adverbs, measures, etc. The third reason is that some properties are implicit and they cannot be extracted by any analyzer.

5 Related Work

Buitelaar et al. [1] is a reference book on ontologies extracted from texts. The different aspects of ontology development are presented: methods, evaluation, and applications. Some approaches aim at building ontologies starting from scratch. For example, Faure et al. [6] use a syntactic structure to describe an object by the verb with which it appears and then statistic measures are used to build a concept hierarchy. Cimiano in [3] use a similar approach but use FCA for building a concept hierarchy. With respect to Cimiano our method proposes a formalization for the resulting ontology, adding defined concepts and we involve knowledge expert.

In the scientific domain, it is important to integrate expert knowledge because some knowledge units are implicit in texts. Stumme et al. [13] merge two ontologies for building a new one. The proposed method takes as input a set of natural language documents. NLP techniques are used to capture two formal contexts encoding the relationships between documents and concepts in each ontology. This method combines the knowledge of the collection of texts and the expert knowledge. In comparaison with our approach, the approach of Stumme et al. uses the texts for merging and not for enriching the two ontologies. Navigli et al. [12] propose to enrich an existing ontology using on-line glossaries. They use natural language definitions of each class and convert them into formal (OWL) definitions, compliant with the core ontology property specifications. Castano et al. [2] also propose to enrich an existing ontology by matching the existing ontology and new knowledge extracted from data. Regarding this methodology, PACTOLE uses a similar idea for evaluating the resulting ontology by similarity between existing and new concepts. This method is called "shallow similarity" by the authors. A difference is that they compare the set of properties while we compare the set of instances.

6 Conclusion and future work

In this paper, we have presented a methodology for semi-automatically enriching an ontology from a collection of texts. This methodology merges a concept hierarchy extracted from a collection of texts with text mining method and a concept hierarchy representing domain knowledge. We have shown how the resulting concept hierarchy can be represented within the DL language $\mathcal{FLE}$. The

proposed methodology was applied to astronomy for extracting knowledge units about celestial objects for problem-solving purposes such as celestial object classification and comparison. We also evaluated the PACTOLE methodology in this context and proposed a definition for precision and recall for evaluating the hierarchy correspondence.

One future work consists in improving the PACTOLE system for the classification of objects annotated "Object of unknown nature" in SIMBAD and suggestion of classes for these objects. Another work consists in integrating relations between the celestial objects in the definition of classes. It is also planned to test the PACTOLE methodology and system in the domain of microbiology domain for the classification of bacteria.

References

1. P. Buitelaar, P. Cimiano, and B. Magnini. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005.
2. S. Castano, A. Ferrara, and G.N. Hess. Discovery-driven ontology evolution. In G. Tummarello, P. Bouquet, and O. Signore, editors, *3rd Italian Semantic Web Workshop*, Pisa, Italy, 2006.
3. P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.
4. M.C. de Marneffe, B. MacCartney, and C.D. Manning. Generating typed dependency parses from phrase structure parses. In *5th International conference on Language Resources and Evaluation (LREC'06)*, GENOA, ITALY, 2006.
5. A. Faatz and R. Steinmetz. Ontology enrichment evaluation. In E. Motta, N. Shadbolt, A. Stutt, and N. Gibbins, editors, *14th International Conference on Engineering Knowledge in the Age of the Semantic Web (EKAW'04)*, volume 3257/2004, pages 497–498, Whittlebury Hall, UK, 2004. Springer.
6. D. Faure and C. Nedellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium. In *11th International Conference in Knowledge Acquisition, Modeling and Management (EKAW'99)*, pages 329–334, Dagstuhl Castle, Germany, 1999. Springer.
7. B. Ganter and R. Wille. *Formal Concept Analysis, Mathematical Foundations*. Springer, 1999.
8. A. Gomez-Perez, M. Fernndez-Lopez, and O. Corcho. *Ontological Engineering*. Springer, 2004.
9. T.R. Gruber. A translation approach to portable ontology specification. *Knowledge Acquisition*, 5:199–220, 1993.
10. Z. Harris. *Mathematical Structure of Language*. Wiley, J. and Sons, 1968.
11. A. Maedche. *Ontology Learning for the Semantic Web*. Springer, 2002.
12. R. Navigli and P. Velardi. Ontology enrichment through automatic semantic annotation of on-line glossaries. In *15th International Conference in Knowlegde Engineering and knowledge Management (EKAW 2006)*, pages 126–140, Czech Republic, 2006. Springer.
13. G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *17th International Joint Conferences on Artificial Intelligence (IJCAI'01)*, pages 225–234, San Francisco, CA, 2001. Morgan Kaufmann Publishers, Inc.
14. A. Valente, T. Russ, R. MacGregor, and W. Swartout. Building and (re)using an ontology of air campaign planning. *IEEE Intelligent Systems*, 14(1):27–36, 1999.