Knowledge Management and Cultural Heritage Repositories

Cross-Lingual Information Retrieval Strategies

Maria Pia di Buono, Mario Monteleone, Federica Marano
University of Salerno
Fisciano (SA), Italy
{mdibuono, mmonteleone, fmarano}@unisa.it

Johanna Monti
University of Sassari
Sassari, Italy
jmonti@uniss.it

***Abstract*— In the last years important initiatives, like the development of the European Library and Europeana, aim to increase the availability of cultural content from various types of providers and institutions. The accessibility to these resources requires the development of environments which allow both to manage multilingual complexity and to preserve the semantic interoperability. The creation of Natural Language Processing (NLP) applications is finalized to the achievement of Cross-Lingual Information Retrieval (CLIR). This paper presents an ongoing research on language processing based on the Lexicon-Grammar (LG) approach with the goal of improving knowledge management in the Cultural Heritage repositories. The proposed framework aims to guarantee interoperability between multi-lingual systems in order to overcome crucial issues like cross-language and cross-collection retrieval. Indeed, the LG methodology tries to overcome the shortcomings of statistical approaches as in Google Translate or Bing by Microsoft concerning Multi-Word Unit (MWU) processing in queries, where the lack of linguistic context represents a serious obstacle to disambiguation. In particular, translations concerning specific domains, as it is has been widely recognized, is unambiguous since the meanings of terms are mono-referential and the type of relation that links a given term to its equivalent in a foreign language is biunivocal, i.e. a one-to-one coupling which causes this relation to be exclusive and reversible. Ontologies are used in CLIR and are considered by several scholars a promising research area to improve the effectiveness of Information Extraction (IE) techniques particularly for technical-domain queries. Therefore, we present a methodological framework which allows to map both the data and the metadata among the language-specific ontologies. This experiment has been set up for the English/Italian language pair and it can be easily extended to other language pairs. The feasibility of cross-language information extraction and semantic search will be tested by implementing an early prototype system.**

***Keywords*— Knowledge Management, Cross-Lingual Information Retrieval, Ontologies.**

I. INTRODUCTION

Building Natural Language Interfaces (NLIs) consists not only in answering questions on the basis of a given database or knowledge base, but also in accessing structured data in the form of ontologies and unstructured data.

The growing need by users to access information on the web in languages different from their own is fostering the research in the field of Cross-language Information Retrieval (CLIR) applications.

Therefore, in this paper, we propose a framework for converting NL queries into formal semantic ones, by means of a procedure which allows to semi-automatically map natural language to formal language. Also, in a more wide perspective, it focuses on the identification of a method for the creation of NLP applications finalized to the achievement of CLIR.

Typically in state-of-the-art CLIR applications, information is searched by means of a query expressed in the user's mother tongue. This query is automatically translated in the desired foreign language and the results are translated back in the user's mother tongue.

This process is based on two different translation stages: query translation and document translation. The query translation concerns the translation in the desired foreign language of the query expressed in the user's mother tongue, whereas the document translation is the back translation in the user's language of the relevant documents found by means of the translated query.

CLIR success obviously depends on the quality of translation and therefore inaccurate translations may cause serious problems in retrieving the relevant information in a foreign language.

The development of an LG-based linguistically motivated system, in which any type of user is able to obtain the exact information he/she is looking for, is not easy to obtain, considering that the first obstacle, not yet solved, is to process a query expressed in natural language. Overcoming such barrier would also mean establishing a method suitable to retrieve information in all languages whether they are formalized or not with reference to their morpho-syntactic features. In this sense, our method aims at being multilingual, thus involving also Machine Translation (MT) techniques and routines.

On the basis of such premises, and starting from a given source language, the system outlined here will follow the following steps:

1. a linguistic analysis inside an NLP environment;

978-1-4799-3170-5/13/$31.00 ©2013 IEEE

2. an iterative transformation finalized to the accomplishment of a machine-readable query;
3. the application of translation routines to obtain multilingual machine-readable queries which translate the original one (multilingual data and meta-data);
4. the execution of such queries against multilingual on-line repositories;
5. the translation of the retrieved results into the source language (i.e. the user's language) and their subsequent display.

It is worth stressing that the processing stages shown in steps 2 and 3 are of crucial importance for reconstructing conceptual relationships among query terms, and also in order to retrieve a meaningful value from subjects' text sequences. As for query words, the matching process to the ontology concepts is also based on domain labels which semantically tag (i.e. denote/connote) each entry of simple-word and multi-word electronic dictionaries[1].

Furthermore, in our NLP environment, finite-state automata (FSA) and finite-state transducers (FSTs) are used to: (a) recognize and classify word relationships inside the query propositions entered/chosen by the user; (b) parse lexical ambiguity.

Indeed, FSA and FSTs are typically applied to locate morpho-syntactic patterns inside corpora and extract matching sequences, in order to build indexes, concordances, etc. This FST/FSA-based method, which is already available inside our NLP environment, can also be used to automatically recognize any kind of text pattern.

In addition, with reference to this environment, an API represents the ideal solution to: (a) build an interface providing procedures callable by means of external processes; (b) drive the application for translating NL queries into Sesame RDF query Language (SeRQL).

A. Background

For several years, we have seen that similar projects and demonstrations proposed data management solutions based on NLIs. Existing proposals share similar goals focusing on the development of applications that meet the required flexibility in order to support the user's view of a given domain. Many of these works have been focused on the use of machine learning algorithms for mapping NL questions to query languages. Indeed, automatic interpretation of natural languages is very difficult to achieve, since the main obstacle in NLIs is the resolution of ambiguity, a problem which is mentioned in various overviews on NLIs [1].

Several design approaches have been used to implement tools which present various levels of expressivity and user-friendliness [2, 3, 4, 5].

There are several approaches to CLIR: they are either based on bilingual or multilingual Machine Readable Dictionaries (MDR), Machine Translation (MT), parallel corpora and finally ontologies.

For a description of the different approaches refer to Hull & Greffenstette [6], Pirkola [7] and more recently Oard [8].

Unlike other NLIs, including some of those previously mentioned, our approach is based on a not statistically-based linguistic formalization which ensures a low degree of ambiguity, a low loss of meaning and an accurate matching between linguistics structures, domain concepts and programming language.

II. METHODOLOGY

Our linguistic methodology is based on the Lexicon-Grammar (LG) theoretical and practical analytical framework. LG theory was conceived by the French linguist Maurice Gross [9], [10] and [11] during the '60s. Unlike Chomsky's transformational grammar and its various offspring [12] and [13], LG assumes that linguistic formal descriptions should be based on the observation of the lexicon and the combinatory behaviours of its elements, encompassing in this way both syntax and lexicon. It has also reached important results in the domain of automatic textual analysis and parsing, with the development of software and lingware fully oriented toward NLP, such as NooJ[2], and former software packages used in the LG framework, such as INTEX and UNITEX[3].

Besides, this methodology is particularly suitable for CLIR applications due to the fact that the quality of translation is guaranteed by a detailed evaluation process[4], thanks to a linguistic approach aimed at the development of a coherent and formalized linguistic knowledge basis. Linguistic Resources (LRs) developed according to the LG framework are used in NLP applications and are helpful to achieve effective Information Retrieval (IR) Systems [14].

In the field of MT-based CLIR, the LG methodology tries to overcome the shortcomings of statistical approaches as in *Google Translate* or *Bing* by Microsoft concerning MWU processing in queries, where the lack of context represent a serious obstacle to disambiguation. LG linguistic framework is grounded in the analysis of the so-called "simple sentence", the smallest linguistic meaning context that can be analysed; on the basis of this "simple sentence" it is possible to achieve concrete studies on natural languages.

The study of simple sentences is achieved by analysing the so-called rules of co-occurrence and selection restriction, i.e. distributional and transformational rules based on predicate syntactic-semantic properties.

Thanks to these abovementioned research studies, LG range of analysis concerns lexicon, and especially the concept of Multiword Unit (MWU) as "meaning unit", "lexical unit" and "word group", for which LG identifies four different combinatorial behaviours [15].

A. Lexicon-Grammar Resources and Tools

As it is well known, LG invests lexicon, and especially the concepts of "meaning unit", "lexical unit" and "word group".

[1] [1] A detailed definition of electronic dictionaries is given in paragraph II letter A.

[2] See http://www.nooj4nlp.net/pages/nooj.html.

[3] More information on the website http://www-igm.univmlv.fr/~unitex/.

[4] For more details see the evaluation methodology par. II letter B.

Besides, it uses FSA/FSTs to retrieve information and parse texts.

In LG MWUs are specific meaning units to be collected in electronic dictionaries. Therefore, we interpret and formalize their formal internal structure by classifying them [16] as Part of Speech patterns[5] (POS) and analysing their semantic properties (Semantic Tagging). Furthermore, we define when a MWU is used compositionally or non-compositionally.

In our electronic dictionaries the morphologic and grammatical characteristics of lexical entries (gender, number and inflection) are formalized by means of distinctive and non-ambiguous alphanumeric tags, which establish ontological relationships between words (entities) and knowledge domains (properties).

Therefore, the development and management of this lexical and ontological database in form of electronic dictionary consist of three main steps [14]:

1. Lexical acquisition. During this on-going phase, MWUs are extracted from corpora and/or certified glossaries and continuously updated.

2. Morpho-grammatical and syntactic tagging. Each lexical entry is given an inflectional paradigm, in order to be inflected.

The following example represents an excerpt extracted from the Italian-English dictionary of Archaeological Artifacts[6]:

freccia di balestra, N+NPN+FLX = C45+DOM = RA1SUOARAL+EN = crossbow arrow,N+NN+FLX=EC3

freccia foliata, N+NA+FLX = C556+DOM = RA1SUOIL+EN = leafed arrow,N+AN+FLX=EC3

fregio con coronamento, N+NPN+FLX = C12+DOM = RA1EDEAES+EN = frieze crown,N+NN+FLX=EC3

fregio dorico, N+NA+FLX = C523+DOM = RA1EDEAES+EN = doric frieze,N+AN+FLX=EC3

fusto a spirale, N+NPN+FLX = C7+DOM = RA1EDEAES+EN = spiral stem,N+AN+FLX=EC3[7]

For each entry, a formal and morphological description is given with:

- the internal structure of each compound. So, in the compound word *fregio dorico* the tag "N" (noun) indicates the grammatical function of the whole compound; the tag "NA" indicates that the given compound is formed by a Noun, followed by an Adjective. At the same time, in the compound word *fregio con coronamento* the tag "NPN", indicates that the given compound is formed by a Noun followed by a Preposition, followed by a Noun;
- the inflectional class. So, the tag "+FLX=C523" indicates the gender and the number of the compound *fregio dorico*, together with its plural form. The inflectional class refers to a local grammar, so, the tag indicates that *fregio dorico* is masculine singular, does not have any feminine correspondent form, and its plural form is *fregi dorici*.

The compound word *fregio dorico* ("Doric frieze") is also marked with the domain tag "DOM=RA1EDEAES", which stands for "Archaeological Artifacts – Building – Architectural Elements – Structural Elements".

The elements that form the morphologic and grammatical patterns of each compound structure are followed by the English translation and its inflectional class.

3. Testing on corpora. The dictionary is used to automatically analyse and process large corpora.

Local grammars are used in NLP routines together with electronic dictionaries. Local grammars are useful to cope with specific characteristics of natural languages; more appropriately, local grammars design is based on syntactic description, which encompasses transformational rules and distributional behaviours [19]. We develop local grammars in the form of FSAs/FSTs [20] and [21].

B. LG Methodology to Assess the Translation Quality

The quality of translations is guaranteed, from the beginning, by developing highly formalized LRs according to morphological, syntactical and semantic criteria. Often using smart translation technologies involves the lowering of Translation Quality (TQ). In LG methodology, instead, we take advantage of well-formed LRs to keep a high level of TQ, since from the beginning, we use a supervised approach carried out by linguists during the proper formalization of resources.

Assessing the quality of resources before they are translated avoids subsequent checks on translated resources, though evaluation ex post of TQ results is necessary in any case.

According to LG a valid evaluation methodology should be based on a hybrid approach that encompasses human and automatic evaluation.

The process, as shown in "Fig. 1", is composed of two cycles. The first cycle can be outlined as follows (i) a query expressed in a Source Language (SL) is the input of the CLIR application, (ii) the CLIR system produces sample queries (i.e. sample texts) in the Target Language (TL), (iii) the resulting translated queries are examined by humans (Linguists, Translators, Terminologists/Domain Experts) to evaluate their quality. The human judgments are based on common criteria of TQ – i.e. adequacy and fluency – and are expressed using a

[5] According to Manning C.D. and Schütze H [17] we consider POS "a part of the grammar of a language which includes the lexical entries for all the words in the language and which may also includes other information".

[6] It's important to specify that our domain dictionaries, collected in the DELAC system, cover about 180 different semantic tags. The most important dictionaries are those of Informatics (54,000 entries ca.), Medicine (46,000 entries ca.), Law (21,000 entries) and Engineering (19,000 entries ca.). Each dictionary has been created and verified under the supervision of domain experts. Subset tags are also previewed for those domains that include specific subsectors. This is the case of Archaeological Artifacts dictionary (9,200 entries ca.), for which a generic tag RA1 is used, while more explicit tags are used for object type, subject, primary material, method of manufacture, object description.

[7] Differently from Italian, the creation of inflection codes for English compound words is still in a preliminary stage. For example, the code EC3 will be used to tag all those compounds which have a structure of type NN / AN, and in which the (second) N inflects like the English word *apple* (plural form: *apples*). At the same time, we stress that up to today 150 inflection codes for English simple names have been created by the linguists of the Maurice Gross Lab for Computational Linguistics (University of Salerno). These codes will be also used to structure and formalize those for English compound words.

Likert scale with scores 1-5 (for instance using the following judgments: 1. Strongly disagree, 2. Disagree, 3. Neither agree nor disagree, 4. Agree, 5. Strongly agree), (iv) only texts which obtained scores 4-5 become "validated" and "supervised" texts which represent the gold standard, (v) this gold standard is the training set for the Automatic Evaluation process, that can be carried out using METEOR[8] and GTM[9], that are the most suitable methods according to our opinion, as well as other ones[10].

During the second cycle, human evaluation is skipped and

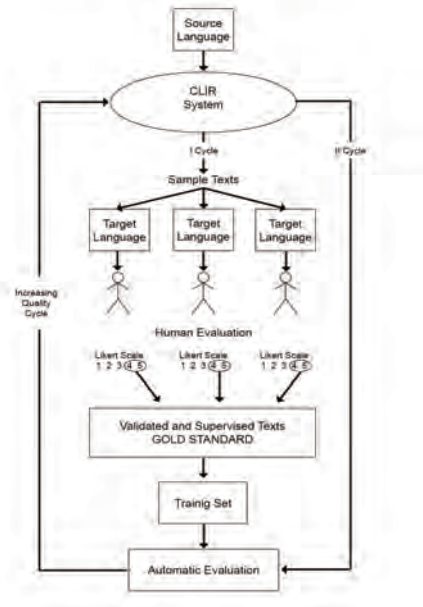

Fig. 1. The hybrid process for Translation Quality Evaluation

the SL queries are directly used as input for automatic evaluation.

It is necessary to periodically repeat the first cycle in order to enrich the training set and to increase the quality cycle.

III. Experiment and Results

Starting from this NLP theoretical and practical framework, in this project we propose to build a User Interface for KMSs which takes as input a NL query from a user, converts it into a SQL query based on domain semantics and database schema, retrieves appropriate data from the database and returns the output to the user. The basic mechanism involves the following iterative transformation:

- the system acquires domain semantics from terminological electronic dictionaries in form of lexical databases;
- it recognizes a NL query by means of local grammars which formalize the query in a linguistic structure;
- it translates the query applying FSA/FSTs;
- it transduces it in an SeRQL path expression.

We propose here an architecture which when applied to a given language, maps data and metadata exploiting the morpho-syntactic and semantic information stored inside both electronic dictionaries and Finite State Automata/Finite State Transducers (FSA/FSTs). In addition, this architecture can map linguistic tags (i.e. POS) and structures (i.e. sentences, MWU) to domain concepts.

The first step performed by our system is a linguistic pre-processing phase which formalizes (i.e. converts) natural language strings into reusable linguistic resources. During this first phase we also extract information from free-form user queries, and match this information with already available ontological domain conceptualizations. As described in "Fig. 2", before the execution of a query against a knowledge base it is necessary to apply the Translation and the Transformation routines. The system is based on two workflows which are carried out simultaneously but independently.

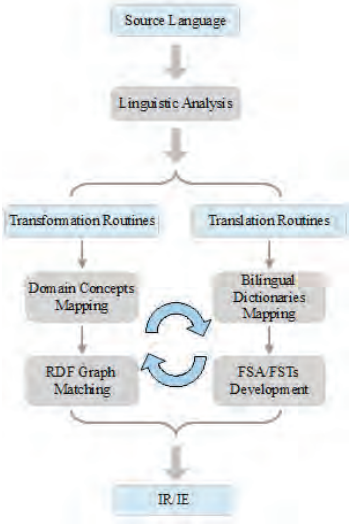

Fig. 2. System Workflow

As shown in "Fig. 2", the process starts with a linguistic analysis based on well-defined Linguistic Resources. Then, transformation Routines are applied to map NL into a RDF-triple graph. Finally, the returned SeRQL query is executed against a knowledge base, in order to extract information and present them to users.

The benefits of keeping separate these two workflows are:

- the development of an architecture with a central multilingual formalization of the lexicon, in which there is no specific target language, but each language can be at the same time target and source language;
- the development of extraction ontologies and SeRQL adaptation systems which could represent a standard not only for our multilingual electronic dictionaries, but also for any lexical and/or language data-base for which translation is required.

[8] http://www.cs.cmu.edu/~alavie/METEOR/.
[9] http://nlp.cs.nyu.edu/GTM/.
[10] BLEU and NIST (based only on precision measure), F-Measure (based also on recall).

Due to all these premises, the described process produces a hybrid architecture, both into the NL analysis and in the base of usable documents. As we will see, NL analysis is hybrid as it copes with strings which are not composed only of words, but also by morpho-grammatical tags (i.e. *N* for any noun, *V* for any verb, and so on). Also, it is hybrid because it may employ three types of information sources, i.e.: (i) unstructured text, (ii) semi-structured and (iii) structured data.

A. Domain modelling and ontology

As for ontologies, the formal definition we rely upon is the one given by the International Council of Museums – Conseil Interational des Musees (ICOM – CIDOC) Conceptual Reference Model (CRM), which defines that "a formal ontology (is) intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" [22]. CIDOC CRM is composed of 90 classes (which include subclasses and superclasses) and 148 unique properties (and sub-properties). The object-oriented semantic model and its terminology are compatible with Resource Description Framework (RDF). Actually, this ontology was already available and is constantly developed. At the same time, our methodology shows that a given linguistic knowledge can be reused independently from the domain to which it pertains. Actually, domain ontologies refer to mid- and upper-level ones, which tend pragmatically to be standardized. Logically, such process indirectly involves also low-level ontologies, and this allows the reuse of linguistic resources regardless of the domain in which they were developed or to which they pertain.

Therefore, LG electronic dictionaries and local grammars may together represent the linguistic (lexical, morphosyntactic and semantic) engine of the KMS [14]. In order to clarify this approach, it becomes necessary to describe the LRs we used to develop our system.

B. Transformation Routine: Transition from NL to RDF Graph

We have seen how a linguistic pre-processing phase may be achieved to formalize natural language strings into reusable linguistic structures. Such structures have the form of knowledge databases, which are transformed into local grammars (FSA/FSTs) for mapping NL query to RDF, and constructing a virtual graph capable to retrieve coherent information.

"Figure 3" shows the process for converting NL text in a SerQL Query. LRs are used for analysing corpora to retrieve recursive phrase structures, in which combinatorial behaviours and co-occurrence between words identify properties, also denoting a relationship. Furthermore, electronic dictionaries also include all inflected verb forms allowing to process queries expressed also with passive and more generally non-

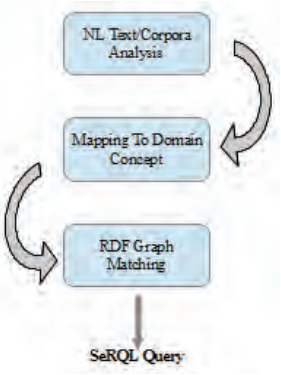

Fig. 3. Transformation Routine

declarative sentences.

Subsequently we use FSA variables for identifying ontological classes and properties for subject, object and predicate within RDF graphs.

This matching of linguistic data to RDF triples and their translation into SERQL path expressions allows the use of specific meaning units to process natural language queries.

"Figure 4" is a sample of an automaton showing an associated RDF graph for the following sentence:

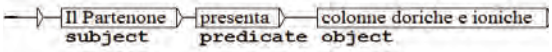

Fig. 4. Simple FSA/FST with RDF Graph

Il Partenone (subject) presenta (predicate) colonne doriche e ioniche (object)

According to our approach, electronic dictionaries entries (simple words and MWUs) are the subject and the object of the RDF triple.

In "Fig. 5" we develop an FSA with a variable which assigns to the sentence the following classes and property:

E19 indicates "Physical Object" class;

P56 stands for "Bears Feature" property;

E26 indicates "Physical Feature" class.

So, the FSA variables transform our sentence into:

Il Partenone (E19) *bears feature* colonne doriche e ioniche (E26).

The role pairs *Physical Object/name* and *Physical Feature/type* are trigged by the RDF predicate *presenta.*

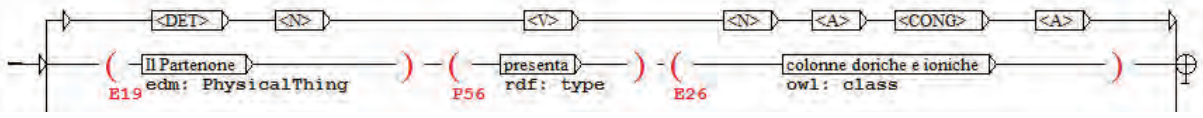

Fig. 5. Sample of the use of the FSA variables for identifying classes for subject, predicate and object.

Besides in "Fig. 5" we also indicate specific POS for the first noun phrase *Il Partenone* (DETerminer + Noun), the verb *presenta* (V) and the second noun phrase *colonne doriche e ioniche* (Noun+Adjective+Conjuntion+Adjective).

By applying the automaton in "Fig. 5" (built using the high variability of lexical classes and not of the original form) we can recognize all instances included in E19 and E26 classes, the property of which is P56.

Indeed, words in angle brackets stand for lemma forms. When the word form is set between angle brackets, the software locates all the word forms that are in the same equivalence set as the given word form (generally all inflected, derived forms, or spelling variants of a given lexical entry).

As we have seen, we choose to use affirmative sentences for mapping linguistic structures to corresponding concepts in the domain ontology. This is due to the fact that generally speaking affirmative sentences are more predictable and reusable from the point of view of word distribution. Therefore, such feature:

- grants a coherent identification and extraction of ontological constrains;
- simplifies the process of information extraction procedure, because it is based on a consistent reusable repository of pre-constituted sentence descriptions.

C. Information extraction

Querying information in a RDF framework means to specify path expressions. Our architecture aims to be useful with the SeRQL query language.

Our specific interest is based on a practical observation: SeRQL uses a path expression syntax which is based on the graph nature of RDF. The path is composed of a collection of nodes and edges and it has an arbitrary length[11].

Indeed, when user queries for two or more triples with identical subject and predicate, the subject and the predicate do not have to be repeated. A multi-value node and branches can be used:

{subj1} pred1 {obj1, obj2, obj3}

This path expression is equivalent to:
{subj1} pred1 {obj1},
{subj1} pred1 {obj2},
{subj1} pred1 {obj3}[12]

This procedure is very close to the linguistic NL features of transformation, deletion and reduction.

In SeRQL we can also apply a restricted form of disjunction through optional matching, and also use existential quantification over predicates and Boolean constraints. Instances of concepts are identified by variables in the subject position of an RDF triple and returns sets of RDF statements.

The query presented in "Fig. 3" can be solved with the following sample (i.e. prototype) path expression:

*SELECT **
FROM
edm:PhysicalThing {PhysicalThing}
owl:ObjectProperty{rdf:about="P33.used_specific_technique"}
rdfs:range {rdf:resource="E29.Design_or_Procedure"}
WHERE
PhysicalThing LIKE "Parthenon"

Where {Production} is a variable representation of Subject; owl:ObjectProperty is the predicate; and rdfs:range is the variable representation of the Object.

D. Translation Routine

We have seen how transformation routines may be applied to convert a NL text in a SeRQL Query. Concerning translation, our methodology is particularly suitable for CLIR applications due to a linguistic approach aimed at the development of a coherent and formalized linguistic knowledge base.

Indeed, a very frequent source of mistranslations in specific domain texts is represented by terminological word compounds. MWUs designate a wide range of lexical constructions, composed of two or more words with an opaque meaning, i.e. the meaning of a unit is not always the result of the sum of the meanings of the single words that are part of the unit. MWUs are not always easy to identify since co-occurrence among the lexemes forming the units may vary a great deal.

Processing and translating these different types of compound words is not an easy task since their morpho-

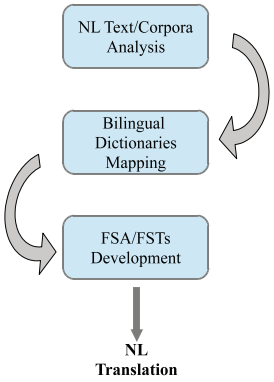

Fig. 6. Translation Routine

syntactic and semantic behaviour is quite complex and varied according to the different types and their translations are practically unpredictable.

"Figure 7" shows a typical Europeana item description in English. The text contains several compound terms (highlighted in the text).

[11] Most current RDF query languages define path expression of length 1 and use them to find combination of triples in an RDF graph.
[12] See Broekstra, Kampman [23].

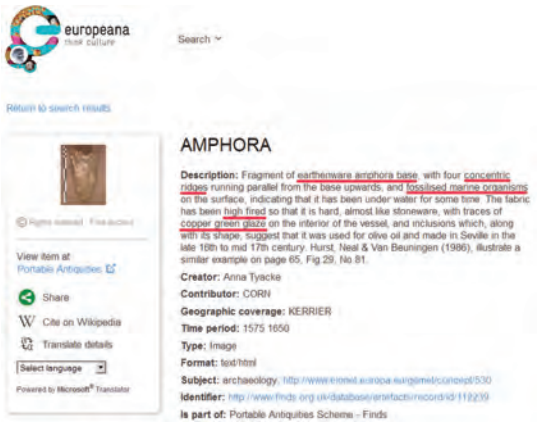

Fig. 7. Sample of Europeana Item Description

"Figure 8" is the result of the automatic translation into Italian of the item description above. Almost all MWU translations powered by Microsoft Translator, the MT system used in Europeana, are wrong, such as *earthenware amphora base* translated with **anfora di terracotta base* instead of *piede di anfora in terracotta* or *high fired* translated with *alto sparato instead of *cotta ad alte temperature.*

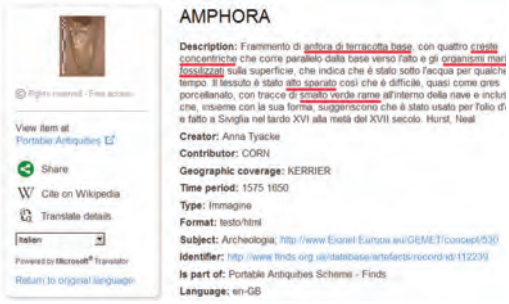

Fig. 8. Sample of Europeana Item Description translated by Microsoft machine translation.

In the field of MT-based CLIR, the LG methodology tries to overcome the shortcomings of statistical approaches as in *Google Translate* or *Bing* by Microsoft concerning MWU processing in queries, where the lack of context represent a serious obstacle to disambiguation. In particular MT-based approaches to CLIR present several limitations in relation to domain-specific contexts (LSPs) where MWUs represent a very frequent and productive linguistic phenomenon. Current approaches to MWU processing in MT move towards the integration of phrase-based models with linguistic knowledge and scholars are starting to use linguistic resources, either hand-crafted dictionaries and grammars or data-driven ones, in order to identify and process MWUs as single units. Monti [24] provides a thorough overview of the problem.

In our approach, an independent and parallel process may be applied to translate the query by means of well-formed LRs and local grammars (FSA/FSTs).

"Figure 9" shows the process for translating NL text in any language. The first phase of this routine is represented by the matching of the results of the Linguistic analysis with bilingual dictionaries. LRs are used to retrieve the correct translation of given linguistic data. As described in Section II, bilingual electronic dictionaries entries include, in addition to the source entry with its morpho-syntactic and semantic information, its equivalent in the target language with the morpho-syntactic and semantic information necessary for a correct translation. In particular, translations concerning specific domains, as it is has been widely recognized, is unambiguous since the meanings of terms are monoreferential and the type of relation that links a given term to its equivalent in a foreign language is biunivocal, i.e. a one-to-one coupling which causes this relation to be exclusive and reversible.

Subsequently we use FSA/FST to apply the transformations necessary to produce the correct NL translation. In our model, the Translation Routines are applied independently of the mapping process of the pivot language. This allows us to preserve the semantic representation in both languages.

Indeed, identifying semantics through FSA/FST guarantees the detection of all data and metadata expressed in any different language.

"Figure 10" shows a FST in which a translation process from Italian to English is performed on the basis of a dictionary look-up, a morpho-syntactic and semantic analysis. This translation FST, in fact, recognizes and annotates the different linguistic elements of declarative sentences such as "Il Partenone presenta fregi dorici", "I templi romani hanno fusti a spirale", etc, with their morpho-syntactic and semantic information and performs automatic translations on the basis of a well-crafted LG bilingual dictionary.

For instance, if a grammar variable, say $E26, holds the value "fusti a spirale", the output $E26$EN will produce the correct translation "spiral stems", on the basis of the value associated to the +EN feature in the bilingual entry " *fusto a spirale,N+NPN+FLX=C7+DOM=RA1EDEAES+EN=spiral stem,N+AN+FLX=EC3"* and the morpho-syntactic analysis performed by the graph in "Fig. 8", which identifies and produces the plural form of the compound noun "fusto a spirale".

IV. Discussion and Conclusions

In this paper, we approach the problem of converting NL queries into programming language.

The proposed architecture ensures not only the coverage of a large knowledge portion but preserves deep semantic relations among different languages.

The aim is to generate metadata representation from natural language inputs. The program outputs RDF graph and SeRQL query representations of a sentences, clauses, and phrases. Furthermore, our architecture ensures a high degree of portability; indeed the specifications are designed to allow the processing of highly complex sentences and phrases of any language and covering any vocabulary.

Future work aims to implement our Linguistic Resources both for testing the accuracy of cross-language information retrieval, extraction and semantic search.

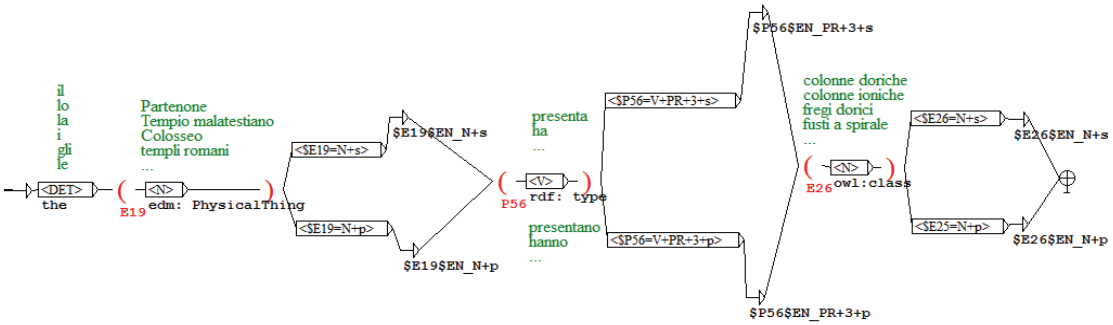

Fig. 9. Sample of the use of the translation FST with variables for identifying classes for subject, predicate and object.

NOTE

Maria Pia di Buono is author of sections III, III.A, III.B and III.C, Johanna Monti is author of sections I, I.A and III.D, Mario Monteleone is author of sections II.A and IV, and Federica Marano is author of section II and II.B.

REFERENCES

[1] Copestake, A. and Sparck Jones K., Natural language interfaces to databases, in: Knowledge Engineering Review, 5(4), *Special Issue on the Applications of Natural Language Processing Techniques*,1989, pp. 225-249.

[2] Lei, Y., Uren, V., Motta E., S*emSearch: a search engine for the semantic web*, in: Managing Knowledge in a World of Networks, Springer Berlin / Heidelberg, 2006, pp. 238-245.

[3] Lopez, V., Motta, E., *Ontology driven question answering in Aqualog*, in: NLDB 2004 (9th International Conference on Applications of Natural Language to Information systems), Manchester, UK, 2004.

[4] Cimiano, P., Haase, P., Heizmann, J., Porting natural language interfaces between domains: an experimental user study with the orakel system, in: IUI '07: *Proceedings of the 12th international conference on Intelligent user interfaces*, New York, ACM, 2007, NY, USA, pp. 180-189.

[5] Kaufmann, E., Bernstein, A., Zumstein, R., Querix: *A natural language interface to query ontologies based on clarification dialogs*, in: 5th International Semantic Web Conference (ISWC 2006), Springer, 2006, pp. 980-981.

[6] Hull D. A., Gregory Grefenstette G. 1996. Querying across languages: a dictionary-based approach to multilingual information retrieval, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval: 49-57.

[7] Pirkola, A. 1998. *The Effects of Query Structire and Dictionary Setups* in Dictionary-Based Cross-language Information Retrieval. In Croft, W.., et al., 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), Melbourne, Australia, August 24-28:55-63.

[8] Oard D. W. 2009. *Multilingual Information Access*. in Encyclopedia of Library and Information Sciences, 3rd Ed., edited by Marcia J. Bates, Editor, and Mary Niles Maack, Associate Editor, Taylor & Francis.

[9] Gross, M. 1968. Grammaire transformationnelle du français. – I – Syntaxe du verbe, Larousse, Paris.

[10] Chomsky, N.A., *Syntactic Structures*. Mouton, The Hague, Paris, 1957.

[11] Gross, M. 1975. Méthodes en syntaxe, régime des constructions complétives, Hermann, Paris.

[12] Gross, M. 1989. La construction de dictionnaires électroniques. Annales des Télécommunications, vol. 44, n° 1-2: 4-19, CENT, Issy-les-Moulineaux/Lannion.

[13] Chomsky, Noam A. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

[14] Marano F., Exploring Formal Models of Linguistic Data Structuring. Enhanced Solutions for Knowledge Management Systems Based on NLP Applications, PhD Dissertation, University of Salerno, Italy, 2012.

[15] De Bueris, G., Elia, A. (eds.): Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche. Plectica, Salerno (2008).

[16] Harris, Z.S. 1970. Papers in Structural and Transformational Linguistics. Reidel, Dordrecht.

[17] Manning C.D. and Schütze H., *Foundations of Statistical Natural Language Processing*, The MIT Press Cambridge, Massachusetts, London, England, 1999.

[18] Vietri S., Elia A., D'Agostino E. 2004. Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian, in Laporte, E., Leclère, C., Piot, M., Silberztein M. (eds.), Syntaxe, Lexique et Lexique-Grammaire. Volume dédié à Maurice Gross, Lingvisticae Investigationes Supplementa 24, John Benjamins, Amsterdam/Philadelphia.

[19] Harris, Z.S. 1957. Co-occurrence and transformation in linguistic structure. Language 33,: 293-340.

[20] Silberztein M. 1993. Dictionnaires électroniques et analyse automatique de textes, Masson, Paris.

[21] Silberztein M. 2002. NooJ Manual. Available for download at: www.nooj4nlp.net.

[22] Crofts, N., Doerr M., Gill T., Stead S., Stiff M., eds., *Definition of the CIDOC Conceptual Reference Model*, Version 5.0, 2008.

[23] Broekstra, J., Kampman A., An RDF Query and Transformation Language, in: *Semantic Web and Peer to Peer*, S. Staab & H. Stuckenschmidt, eds., Springer Berlin Heidelberg, Berlin, 2006, pp 23-29.

[24] Monti, J. Multi-word unit processing in Machine Translation: developing and using language resources for multi-word unit processing in Machine Translation. PhD dissertation. University of Salerno, Italy, 2013.