Multi-word processing in an ontology-based Cross-Language Information Retrieval model for specific domain collections

Maria Pia di Buono
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
mdibuono@unisa.it

Johanna Monti
UNISS
Via Roma 151
Sassari, Italy
jmonti@uniss.it

Mario Monteleone
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
mmonteleone@unisa.it

Federica Marano
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
fmarano@unisa.it

Abstract

This paper proposes a methodological approach to CLIR applications for the development of a system which improves multi-word processing when specific domain translation is required. The system is based on a multilingual ontology, which can improve both translation and retrieval accuracy and effectiveness. The proposed framework allows mapping data and metadata among language-specific ontologies in the Cultural Heritage (CH) domain. The accessibility of Cultural Heritage resources, as foreseen by recent important initiatives like the European Library and Europeana, is closely related to the development of environments which enable the management of multilingual complexity. Interoperability between multilingual systems can be achieved only by means of an accurate multi-word processing, which leads to a more effective information extraction and semantic search and an improved translation quality.

1 Introduction

Cross-language Information Retrieval (CLIR) applications aimed at accessing information on the web in several languages is attracting many important players in the Information Retrieval (IR) field, such as Google and Microsoft. Typically in 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. Translation is usually based on bilingual or multilingual Machine Readable Dictionaries (MRD), Machine Translation (MT) and parallel corpora.

CLIR applications are often used in domain specific collections, such as the Europeana Connect, which is aimed at facilitating multilingual access to Europeana.eu, an internet portal that acts as an interface to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe, regardless of the users' native language.

In Europeana Connect, indeed, users can submit queries in their native language and are able to retrieve documents in other languages and ob-

Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.
Monti, J., Mitkov, R., Corpas Pastor, G., Seretan. V., eds.

tain information about objects from many sources across all European countries. The retrieved information is translated back into the user's language by means of MT.

Figure 1 shows a typical Europeana item description in English. The text contains several compound terms (highlighted in the text). Compound terms belong to multi-word units (MWU), which designate a wide gamut 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. In domain specific texts compound terms, mainly noun compounds, are very frequent. In all languages there is indeed a close relationship between terminology and multi-words and, in particular, word compounds. In fact, word compounds account in some cases for 90% of the terms belonging to a domain specific language.

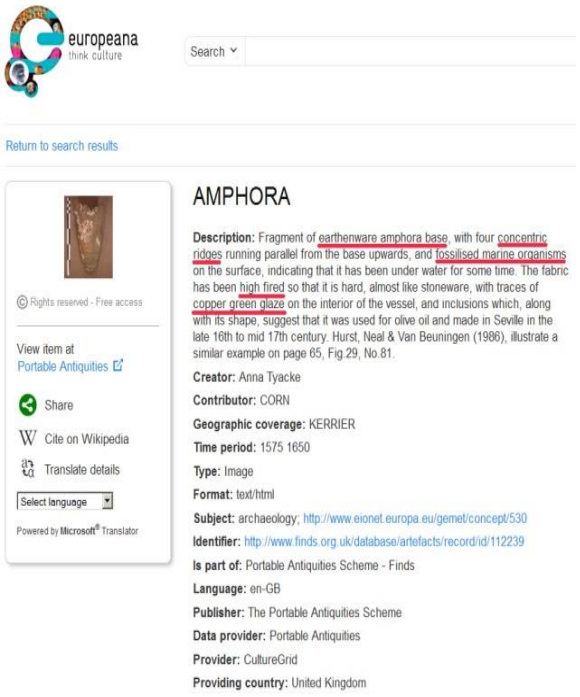

Figure 1: Europeana item description

CLIR success clearly depends on the quality of translation and therefore inaccurate or incorrect translations may cause serious problems in retrieving relevant information. A very frequent source of mistranslations in specific domain texts, as clearly emerges from the example in Figure 2, is, indeed, represented by MWUs, and in particular terminological word compounds.

Contrary to generic simple words, terminological word compounds are mono-referential, i.e. they are unambiguous and refer only to one specific concept in one special language, even if they may occur in more than one domain. Their meaning, similar to all compound words, cannot be directly inferred by a non-expert from the different elements of the compounds because it depends on the specific area and the concept it refers to.

Figure 2 is the result of the automatic translation into Italian of the item description in Figure 1. 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.*

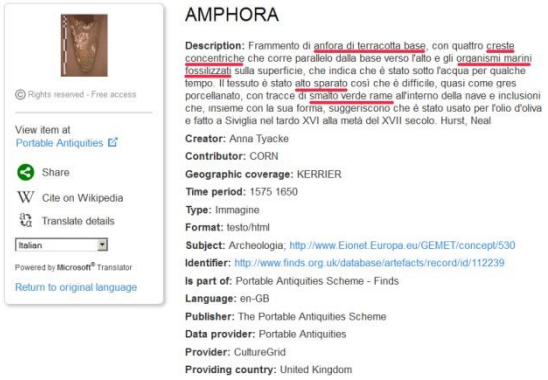

Figure 2: Europeana item description translated by Microsoft Machine Translation

Processing and translating these different types of compound words is not an easy task since their morpho-syntactic and semantic behavior is quite complex and varied according to the different types and their translations are practically unpredictable.

The main contribution of this paper is the experimentation of an ontology-based CLIR system designed to overcome the current limitations of the state-of the-art CLIR, in specific domain collections, and in particular to take into account a proper processing and translation of MWUs. This experiment has been set up for the Italian/English

language pair and can be easily extended to other language pairs.

The remaining of this paper is organized as follows. The next section briefly explains the related work in the area of CLIR. Section 3 describes the methodology used in the experiment. Then, section 4 is devoted to system overview, and, in particular, presents the data modeling and the system architecture extension. Section 5 introduces the feasibility study together with the description of the electronic dictionaries, the semantic annotation and the translation process. Finally, conclusions and future work are described in section 6.

2 Related work

Approaches to CLIR are either based on bilingual or multilingual Machine Readable Dictionaries (MRD), Machine Translation (MT), parallel corpora and finally ontologies.

Hull & Greffenstette (1996), Oard & Dorr (1996), Pirkola (1998) and more recently Oard (2009) provide comprehensive descriptions of these approaches.

Both MRD-based and MT-based CLIR are the prevalent models but they show several weaknesses especially with regard to domain-specific contexts because they are not able to solve translation problems associated to MWUs, a very frequent and productive linguistic phenomenon in languages for special purposes (LSPs). Both approaches in most cases produce literal translations of the single constituents of MWUs which do not represent appropriate translation solutions for this type of lexical constructions. MWUs, in fact, have to be considered as single meaning units. For instance, the Italian translation of the compound adjective "*high fired"* is *cotto ad alte temperature* which cannot be obtained by the literal translation of the single constituents of this MWU.

Translation errors mainly depend on lack of coverage and quality of the systems and various techniques have been proposed to reduce the errors due to the presence of MWU used during query translation. Among these techniques, phrasal translation, co-occurrence analysis, and query expansion are the most popular ones.

Concerning phrasal translation, techniques are often used to identify multi-word concepts in the query and translate them as phrases. Hull & Grefenstette (1996) showed that the performance achieved by manually translating phrases in queries is significantly better than that of a word-by-word translation using a dictionary. Davis and Ogden (1997) used a phrase dictionary extracted from parallel sentences in French and English to improve the performance of CLIR. Ballesteros and Croft (1996) performed phrase translation using information on phrase and word usage contained in Collins MRD. More recently, Gao et al. (2001) propose that noun phrases are recognized and translated as a whole by using statistical models and phrase translation patterns and that the best word translations are selected based on the cohesion of the translation words. Finally, Saralegi & de Lacalle (2010) use a simple matching and translation technique based on a bilingual MWU list to detect and translate them.

Co-occurrence statistics is used to identify the best translation(s) among all translation candidates using text collections in the target language as a language model, assuming that correct translations occur more frequently than wrong ones (Maeda et al., 2000; Ballesteros and Croft, 1998; Gao et al., 2001; Sadat et al., 2001).

As for query expansion techniques, Ballesteros & Croft (1996 and 1997) assume that additional terms that are related to the primary concepts in the query are likely to be relevant and that phrases in query expansion via local context analysis and local feedback can be used to reduce the error associated with automatic dictionary translation.

Concerning MT-based CLIR, MWU identification and translation problems are far from being solved. MWU processing and translation in SMT started being addressed only very recently and different solutions have been proposed so far, but basically they are considered either as a problem of automatically learning and integrating translations, of word alignment or word sense disambiguation (WSD) (Monti, 2013).

Current approaches to MWU processing move towards the integration of phrase-based models with linguistic knowledge and scholars are starting to use linguistic resources (LRs), either handcrafted dictionaries and grammars or data-driven ones, in order to identify and process MWUs as single units.

A first possible solution is the incorporation of MRDs and glossaries into the SMT system, for which there are several straightforward approaches. One is to introduce the lexicon as phrases in the phrase-based table. Unfortunately, the words coming from the dictionary have no

context information. A similar approach is to introduce them to substitute the unknown words in the translation, but this poses the same problem as before.

Another solution for overcoming translation problems in MT and in SMT in particular is based on the idea that MWUs should be identified and bilingual MWUs should be grouped prior to statistical alignment (Lambert and Banchs, 2006). In their work, bilingual MWU were grouped as one unique token before training alignment models.

More recently, Ren et al. (2009) have underlined that experiments show that the integration of bilingual domain MWUs in SMT could significantly improve translation performance. Wu et al. (2008) propose the construction of phrase tables using a manually-made translation dictionary in order to improve SMT performance. Finally, Bouamor et al. (2011) affirm that integration of contiguous MWUs and their translations improves SMT quality and propose a hybrid approach for extracting contiguous MWUs and their translations in a parallel corpus.

Other solutions try to integrate syntactic and semantic structures (Chiang, 2005; Marcu et al., 2006; Zollmann & Venugopal, 2006), but the solutions undoubtedly vary according to the different degrees of compositionality of the MWU.

Very recently, identification and disambiguation of MWUs are being considered as a problem of Word Sense Disambiguation (WSD), i.e. the identification and the selection of the proper meaning of a word in a given context when it has multiple meanings, and several approaches to integrate WSD in SMT have been proposed (Carpuat & Wu, 2007; Carpuat & Diab, 2010 among others).

The problem is here to select the most appropriate translation in TL to a given lexical unit in the SL. Some scholars refer to this problem also as word translation disambiguation (WTD), such as for instance Yang and Kirchoff (2012).

Ontologies are also 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. Volk et al. (2003) use ontologies as interlingua in cross-language information retrieval in the medical domain and show that the semantic annotation outperforms machine translation of the queries, but the best results are achieved by combining a similarity thesaurus with the semantic codes. Yapomo et al. (2012) perform ontology-based query expansion of the most relevant terms exploiting the synonymy relation in WordNet.

3 Methodology

Our approach to CLIR is based on Lexicon-Grammar (LG) devised by the French linguist Maurice Gross during the ‘60s (Gross, 1968, 1975 and 1989).

LG presupposes that linguistic formal descriptions should be based on the examination of the lexicon and the combinatory behaviors of its elements, encompassing in this way both syntax and lexicon. Nowadays, the LG methodology is being adopted by a wide research community both for Indo-European languages (French, Italian, Portuguese, Spanish, English, German, Norwegian, Polish, Czech, Russian, Bulgarian and Greek) and other ones (Arabic, Korean, Malagasy, Chinese, Thai...).

LG linguistic framework is based on the analysis of the so-called “simple sentence” [1], the smallest linguistic meaning context, by applying rules of co-occurrence and selection restriction.

LG scholars have been studying MWUs for years now and LG research in this field is indebted to the transformational and distributional concepts developed by Harris (1957, 1964 and 1982).

Thanks to these abovementioned research studies, LG range of analysis concerns lexicon, and especially the concept of MWU as “meaning unit”, “lexical unit” and “word group”, for which LG identifies four different combinatorial behaviors (De Bueriis and Elia, 2008).

Linguistic resources (LRs) developed according to the LG framework are used in Natural Language Processing (NLP) applications and are useful to achieve effective Information Retrieval (IR) systems (Marano F., 2012) and translation processes.

In the field of CLIR, the LRs developed according to the LG methodology can be used to overcome the shortcomings of statistical approaches to MT such as in *Google Translate* or

[1] In LG, a simple sentence is a context formed by a unique predicative element (a verb, but also a name or an adjective) and all the necessary arguments selected by the predicate in order to obtain an acceptable and grammatical sentence. For a detailed definition of simple sentence refer to Gross (1968).

Bing by Microsoft concerning MWU processing in queries, where the lack of context represent a serious obstacle to disambiguation. The same resources can also be used for domain-adaptation purposes in SMT, thus improving the translation quality in the document translation phase in specific domain contexts.

The main linguistic resources developed by LG researchers concerning MWUs are (i) matrix tables describing the syntactic-semantic properties of lexical entries, (ii) morphologically and semantically tagged electronic dictionaries, (iii) local grammars in the form of Finite State Automata (FSA) [2] and Finite State Transducers (FST)[3].

3.1 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 deterioration 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 highly skilled linguists during the proper setting of the resources.

Assessing the quality of resources before they are translated prevents from 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 both human and automatic evaluation.

The process 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 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 obtain 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[4] and GTM[5], the most suitable methods according to our opinion, as well as other ones[6].

During the second cycle, human evaluation is skipped and 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.

4 System overview

We propose an architecture, which, when applied to a given language, maps data and metadata exploiting the morpho-syntactic and semantic information stored both in electronic dictionaries and FSA/FSTs (presented in 5.2 and 5.3). Furthermore, this architecture can also 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. 3, prior to the execution of a query against a knowledge base it is necessary to apply the Translation and the Transformation routines. We can see that the system is based on two workflows which are carried out simultaneously but independently.

[2] Finite-State Automata (FSA) are a special case of Finite-State Transducers that do not produce any result (i.e. they have no output). Typically, FSA are used to locate morpho-syntactic patterns in corpora and extract the matching sequences to build indices, concordances, etc.

[3] Finite-State Transducers (FSTs) are graphs that represent a set of text sequences and then associate each recognized sequence with an analysis result. The text sequences are described in the input part of the FST; the corresponding results are described in the output part of the FST. Typically, a syntactic FST represents word sequences and then produces linguistic information (its phrasal structure, for example).

[4] http://www.cs.cmu.edu/~alavie/METEOR/.

[5] http://nlp.cs.nyu.edu/GTM/.

[6] BLEU and NIST (based only on precision measure), F-Measure (based also on recall).

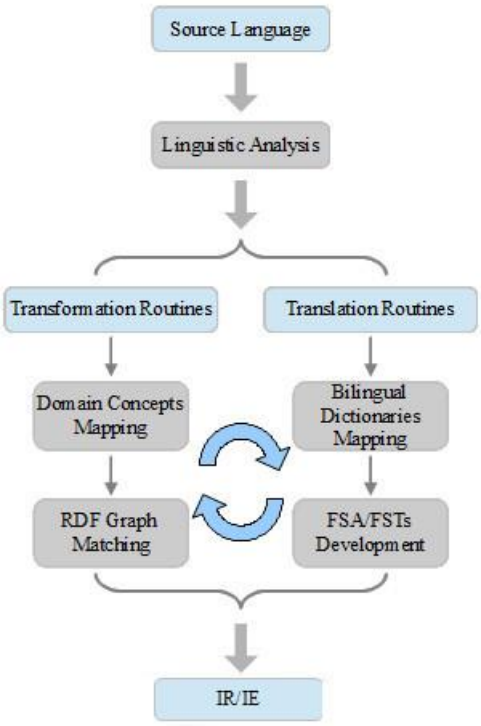

Figure 3: System workflow

The benefits of keeping separate these two workflows are (i) 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, (ii) the development of extraction ontologies and SPARQL/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.

With this dual-structure system, it is easier to successfully achieve the CLIR process since the results are given explicitly in the target language chosen by the user and the translation process is separated from the matching with the RDF triples.

5 Feasibility study

To test the feasibility of our architecture, we are carrying out a translation experiment from Italian into English, using all ontological and semantic constraints defined for the Italian model.

We have chosen the Archaeological domain to test the applicability of our approach. This choice allows us to demonstrate that the modularity of our architecture may be applied to a domain which is variable by type and properties and is semantically interlinked with other domains.
In the next paragraphs, we will present the LRs developed for our study, together with the description of the semantic annotation and the translation routines used in query translation.

5.1 Electronic dictionaries

An electronic dictionary is a lexical database homogeneously structured, in which the morphologic and grammatical characteristics of lexical entries (gender, number and inflection) are formalized by means of distinctive and non-ambiguous alphanumeric tags (Vietri et al. 2004).

All the electronic dictionaries, developed according to the LG descriptive method, form the DELA[7] system, which is used as the linguistic knowledge base in NLP applications. DELA electronic dictionaries are of two types: (i) simple word dictionaries, which include semantically autonomous lexical units formed by character sequences, delimited by blanks, such as *home* and *chair*, (ii) compound word dictionaries, which include lexical units composed of two or more simple words with a non-compositional meaning, such as *nursing home* and *rocking chair*. Terminological compound words (the most common obstacle in CLIR applications) are lemmatized in compound word electronic dictionaries[8].

The following example represents an excerpt from the Italian/English compound word dictionary of Archaeological Artefacts:

anfora di terracotta, N + NPN + FLX=C41 + DOM=RA1 + EN=earthenware amphora, N+AN+FLX=EC3
cerchi concentrici, N + NA + FLX=C601 + DOM=RA1 + EN=concentric ridges, N+AN+FLX=EC4

[7] Dictionnaire Électronique of LADL (Laboratoire d'Automatique Documentaire et Linguistique).

[8] Our domain dictionaries 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.). Subset tags are also foreseen for those domains that include specific subsectors. This is the case of Archaeological Artefacts 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.

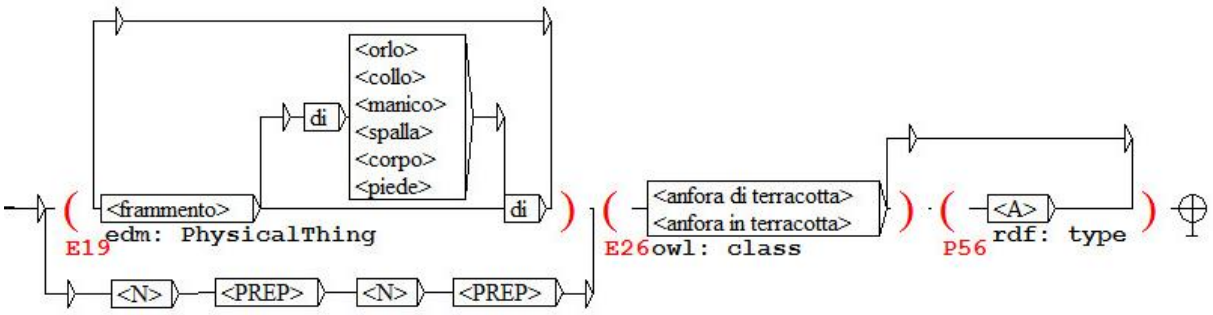

Figure 4: Use of FSA variables for identifying classes for subject, predicate and object

cottura ad alte temperature, N + NPAN + FLX =C611 + DOM=RA1 + EN=high fired, N+AN+FLX=EC4
fregio dorico, N + NA + FLX = C523 + DOM=RA1 + EN=doric frieze, N+AN+FLX=EC3
fusto a spirale, N + NPN + FLX = C7 + DOM=RA1 + EN=spiral stem, N+AN+FLX=EC3
fossile marino, N + NA + FLX = C501 + DOM=RA1 + EN=fossilised marine organism, N+AN+FLX=EC3
smalto verde rame, N + NAN + FLX=C04 + DOM=RA1 + EN=copper green glaze, N+AN+FLX=EC4

The compound words belong to the «Archaeological Artifacts» domain, marked with the domain tag «DOM=RA1» in the dictionary.

For each entry, a formal and morphological description is also given with (i) the internal structure of each compound, such as in the compound word *fregio dorico,* where the tag «NA» specifies that it is formed by a Noun, followed by an Adjective. (ii) the inflectional class, such as the tag «+FLX=C523», which indicates the gender and the number of the compound *fregio dorico,* together with its plural form, i.e. that *fregio dorico* is masculine singular, does not have any feminine corresponding form, and its plural form is *fregi dorici*. Each inflection class is associated to a local grammar which produces all the inflected forms of the compound words according to the inflection class associated to them.

Together with electronic dictionaries, local grammars are used in NLP routines to parse texts. Local grammars are useful to cope with specific characteristics of natural language; more appropriately, local grammars design is based on syntactic descriptions, which encompasses both transformational rules and distributional behaviours (Harris, 1957). Local grammars are developed in the form of FSA/FST (Silberztein, 1993 and 2002)[9].

5.2 Semantic annotation

As for ontologies, the formal definition we rely upon is the one given by the International Council of Museums - Conseil International des Musèes (ICOM – CIDOC) Conceptual Reference Model (CRM), which states that "a formal ontology (is) intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" (Crofts et al., 2008).

CIDOC CRM is a core ontology composed of 90 classes (which includes subclasses and superclasses) and 148 unique properties (and subproperties). The object-oriented semantic model and its terminology are compatible with the Resource Description Framework (RDF). This ontology is constantly developed and updated.

We use FSA variables for identifying ontological classes and properties for subject, object and predicate within RDF graphs, as presented in Figure 4. FSA are based on LRs, which are used during the analysis of corpora to retrieve recursive phrase structures, in which combinatorial behaviours and co-occurrence between words identify properties, also denoting a relationship. Furthermore, electronic dictionaries include all inflected verb forms allowing to process queries

9 To develop and test electronic dictionaries and local grammars we use the NooJ software, an NLP environment, based on the DELA system of electronic dictionaries, on LG syntactic tables and on FSA/FST, developed in the form of graphs and used in LG to parse texts.

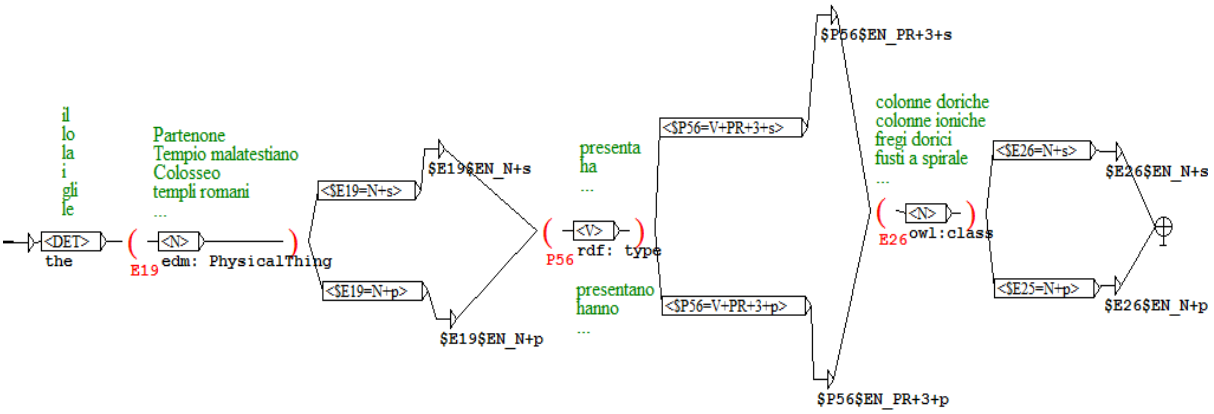

Figure 5: Example of a translation FST

expressed also with passive and more generally non-declarative sentences.

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

Figure 4 is a sample of an automaton which recognizes the following MWU:

> *frammento di (*Empty + *orlo* + *collo* + *manico* + *spalla* + *corpo* + *piede) (*Empty + *di) (anfora di terracotta* + *anfora in terracotta)* (Empty + any adjective)

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

In Figure 4 we also use FSA variables which apply to the sentence the following CIDOC-CRM classes and property: (i) E19 indicates "Physical Object" class; (ii) P56 stands for "Bears Feature" property; (iii) E26 indicates "Physical Feature" class.

Together with FSA variables we also associate POS to the Europeana Semantic Elements (ESE) metadata format[10], currently used in Europeana, i.e. edm: PhysicalThing, owl: class, rdf: type.

Furthermore, the automaton, built using lexical classes (Fig. 4), recognizes all instances included in E19 and and E26 classes, the property of which is P56, and not only the original MWUs.

5.3 Query 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 guarantees the detection of all data and metadata expressed in any different language.

Figure 5 shows an 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, identifies 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 an 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 Figure 5, which identifies and produces the plural form of the compound noun "fusto a spirale".

6 Conclusions and future work

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

Future work aims at implementing our Linguistic Resources to test the accuracy of cross-

[10] http://pro.europeana.eu/edm-documentation

language information retrieval, extraction and semantic search.

Note

Maria Pia di Buono is author of sections 4, 5 and 5.2, Johanna Monti is author of sections 1, 2 and 5.3, Mario Monteleone is author of sections 5.1 and 6 and Federica Marano is author of section 3 and 3.1.

References

Ballesteros L. and Croft B. 1996. *Dictionary Methods for Cross-Lingual Information Retrieva*l. Proc. of the 7th DEXA Conference on Database and Expert Systems Applications, Zurich, Switzerland, September 1996: 791-801.

Ballesteros L. and Croft B. 1997. *Phrasal translation and query expansion techniques for crosslanguage information retrieval*. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval.

Ballesteros L. and Croft B. 1998. *Resolving Ambiguity for Cross-language Retrieval*. SIGIR'98, Melbourne, Australia, August 1998: 64-71.

Bouamor D., Semmar N., and Zweigenbaum, P. 2011. *Improved statistical machine translation using multi-word expressions.* Proceedings of MT-LIHMT. Barcelona, Spain.

Carpuat M. and Diab M. 2010. *Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation*. HLT-NAACL 2010.

Carpuat M. and Wu D. 2007. *Improving statistical machine translation using word sense disambiguation.* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL): 61-72.

Chiang, D. 2005. *A hierarchical phrase-based model for statistical machine translation.* Proceedings of Association of Computational Linguistics (ACL).

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

Davis M. W., and Ogden W. C. 1997. *Free resources and advanced alignment for cross-language text retrieval.* The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersbury, MD.

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

Gao J., Nie J., Xun E., Zhang J., Zhou M., Huang C. 2001. *Improving Query Translation for Cross-Language Information Retrieval using Statistical Models.* Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.

Gross M. 1968. *Grammaire transformationnelle du français. – I – Syntaxe du verbe*, Larousse, Paris.

Gross M. 1975. *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris.

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

Harris Z.S. 1957. *Co-occurrence and transformation in linguistic structure*. Language 33: 293-340.

Harris Z.S. 1964. *Transformations in Linguistic Structure.* Proceedings of the American Philosophical Society 108:5:418-122.

Harris Z.S. 1982. *A Grammar of English on Mathematical Principles*. John Wiley and Sons, New York, USA.

Hull D. A. and 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.

Lambert P. and Banchs R. 2006. *Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation*. Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context. Trento, Italy.

Maeda, A., Sadat, F., et al. 2000. *Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine*. Proc. of the Fifth Int'l Workshop on Info. Retrieval with Asian Languages, Hong Kong, China: 173-179.

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

Marcu D., Wei W., Echihabi A., and Knight K. 2006. *SPMT: Statistical Machine Translation with Syntactified Target Language Phrases*. Proceedings of Empirical Methods in Natural Language Processing (EMNLP).

Monti, J. 2013. *Multi-word unit processing in Machine Translation: developing and using language resources for multi-word unit processing in Ma-*

chine Translation. PhD dissertation. University of Salerno, Italy.

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

Oard, D. W. and Dorr, B. J. 1996. *A survey of multilingual text retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Pirkola A. 1998. *The Effects of Query Structure 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.

Ren, Z. Lü, Y., Cao J., Liu Q., and Zhixiang Y. 2009. *Improving statistical machine translation using domain bilingual multiword expressions*. Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore : 47-54.

Sadat F., Maeda A., Yoshikawa M. and Uemura S. 2001. *Query expansion techniques for the CLEF bilingual track*. Working Notes for the CLEF 2001 Workshop: 99-104.

Saralegi X. and de Lacalle M. L. 2010. *Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR*. Proceedings of the 7th International Conference on Language Resources and Evaluations (LREC). Malta.

Silberztein M. 1993. *Dictionnaires électroniques et analyse automatique de textes*, Masson, Paris.

Silberztein M. 2002. *NooJ Manual*. Available for download at: www.nooj4nlp.net.

Szpektor I., Dagan I., Lavie A., Shacham D., Wintner S. 2007. *Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation*. Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data, Prague, Czech Republic.

Vietri S., Elia A. and D'Agostino E. 2004. *Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian*, 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.

Volk M., Vintar S., and Buitelaar P. 2003. *Ontologies in cross-language information retrieval.* Proceedings of WOW2003 (Workshop Ontologie-basiertes Wissensmanagement), Luzern, Switzerland.

Vossen P., Soroa A., Zapirain B. and Rigau G. 2012. *Cross-lingual event-mining using wordnet as a shared knowledge interface.* Proceedings of the 6th Global Wordnet Conference, C. Fellbaum, P. Vossen (Eds.), Publ. Tribun EU, Brno, Matsue, Japan, January 9-13:382-390.

Wu, H., Wang, H., & Zong, C. 2008. *Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora.* Proceedings of Conference on Computational Linguistics (COLING): 993-100.

Yang M. and Kirchhoff K. 2012.*Unsupervised Translation Disambiguation for Cross-Domain Statistical Machine Translation*. Proceedings of AMTA.

Yapomo M., Corpas G. and Mitkov R. 2012. *CLIR- and ontology-based approach for bilingual extraction of comparable documents*. The 5th Workshop on Building and Using Comparable Corpora.

Zollmann A., and Venugopal A. 2006. *Syntax augmented machine translation via chart parsing*. In Proceedings of the Workshop on Statistical Machine Translation, HLT/NAACL.