Multi-word processing in domain specific Cross-Language Information Retrieval applications: integration of ontology-based methods and Linguistic resources

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

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

Abstract

This poster proposes a methodological approach to Cross-language Information Retrieval (CLIR) applications for the development of a system which improves multi-word (MWE) 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 CH 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.

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.

CLIR success clearly depends on the quality of translation and therefore inaccurate or incorrect translations may cause serious problems in retrieving relevant information.

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.

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

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. Furthermore, this architecture can also map linguistic tags (i.e. POS) and structures (i.e. sentences, MWE) 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. Prior to the execution of a query against a knowledge base it is necessary to apply the Translation and the Transformation routines.

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.

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

Anfora di terracotta,N+NPN+FLX=C41+DOM=RA1+EN= earthenware amphora,N+AN+FLX=E3
cerchi concentrici,N+NA+FLX=C601+DOM=RA1+EN= concentric ridges,N+AN+FLX=EC4
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

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

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). The object-oriented semantic model and its terminology are compatible with the Resource Description Framework (RDF).

We use FSA variables for identifying ontological classes and properties for subject, object and predicate within RDF graphs. According to our approach, electronic dictionaries entries (simple words and MWEs) are the subject and the object of the RDF triple. 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, currently used in Europeana, i.e. edm: PhysicalThing, owl: class, rdf: type. Furthermore, the automaton, built using lexical classes, recognizes all instances included in E19 and E26 classes, the property of which is P56, and not only the original MWEs.

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. The translation process from Italian to English is performed on the basis of a dictionary look-up, a morpho-syntactic and semantic analysis. 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, which identifies and produces the plural form of the compound noun "fusto a spirale".

References

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.