ECP-2008-DILI-528001

EuropeanaConnect

D2.4.1 – Modules for Partial Query and Result Translation / Enrichment

Deliverable number/name	*D2.4.1*
Dissemination level	*Public*
Delivery date	*30/04/2011*
Status	*Final*
Author(s)	*Alessio Bosca*

e*Content*plus

This project is funded under the *e*Content*plus* programme,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

Österreichische Nationalbibliothek EuropeanaConnect is coordinated by the Austrian National Library

D2.4.1 – Modules for Partial Query and Result Translation / Enrichment

co-funded by the European Union

The project is co-funded by the European Union, through the ***eContentplus*** programme

http://ec.europa.eu/econtentplus

Österreichische
Nationalbibliothek

EuropeanaConnect is coordinated by the Austrian National Library

Table of Contents

1 Introduction

EuropeanaConnect work package 2 aims at providing multilingual access for international users. One major task is to set up translation modules or services for cross-lingual user queries. A proper software module named MultiLingual Information Access (MLIA) has been designed and included in the Language Resource Repository (LRR), described in Deliverable D2.2.1 – Europeana Language Resources Repository and appropriate description & licensing mechanism. The MLIA module is located at the top level of the LRR in the directory named **mlia** (see Section 2 of Deliverable D2.2.1, "*Accessing and using the repository*").

The goal of the MLIA module consists in providing query translation functionalities to the Europeana portal in order to support Cross-Language Information Retrieval strategies. The MLIA module implements a query translation strategy by exploiting and coordinating the Language Resources present in the Europeana Language Resources Repository. The queries input by users is enriched with linguistic and semantic annotations (i.e. part of speech, named entity category) and then translated into target languages via bilingual dictionaries. The MLIA specifically depends on the following types of resources (detailed in Section 4.2 of Deliverable D2.2.1, "*Resource Description*"):

- **Language Identifiers**: a tool that is necessary whenever the language of the query is not explicitly known.
- **Morphological Analyzers**: software modules that perform tokenization and lemmatization, but also decompounding, multi-word detection and Part of Speech tagging.
- **Named Entity Recognizers**: software modules that identify named entities, such as person names, geographic names, organisation names, etc.
- **Translation Dictionaries**: mappings between terms in different languages.

This document describes the MLIA module reporting the translation strategy implemented by the module and the programmatic interface exposed by the module to the portal. **Section 2** describes the approach used by the module and the interactions between the MLIA module and the language resources needed. **Section 3** presents the software interface used by the module along with an example usage that shows how to programmatically access it and the available options for MLIA configuration. **Section 4** shows a few examples of translations performed by the module.

Within this document a few examples of input queries are reported in order to illustrate the approach adopted in query translation as well as the obtained results (see section 3 and 4 of the present document). Such examples are taken from the ongoing tests on the translation modules that are part of task 2.5 (**Sandbox integration, testing and evaluation of translation modules**). The evaluation tests on the translation modules are still ongoing and will be described in details in the Milestone document M2.5.1 (Implementation of modules in Europeana sandbox environment and evaluation of translation modules (document, month 26))

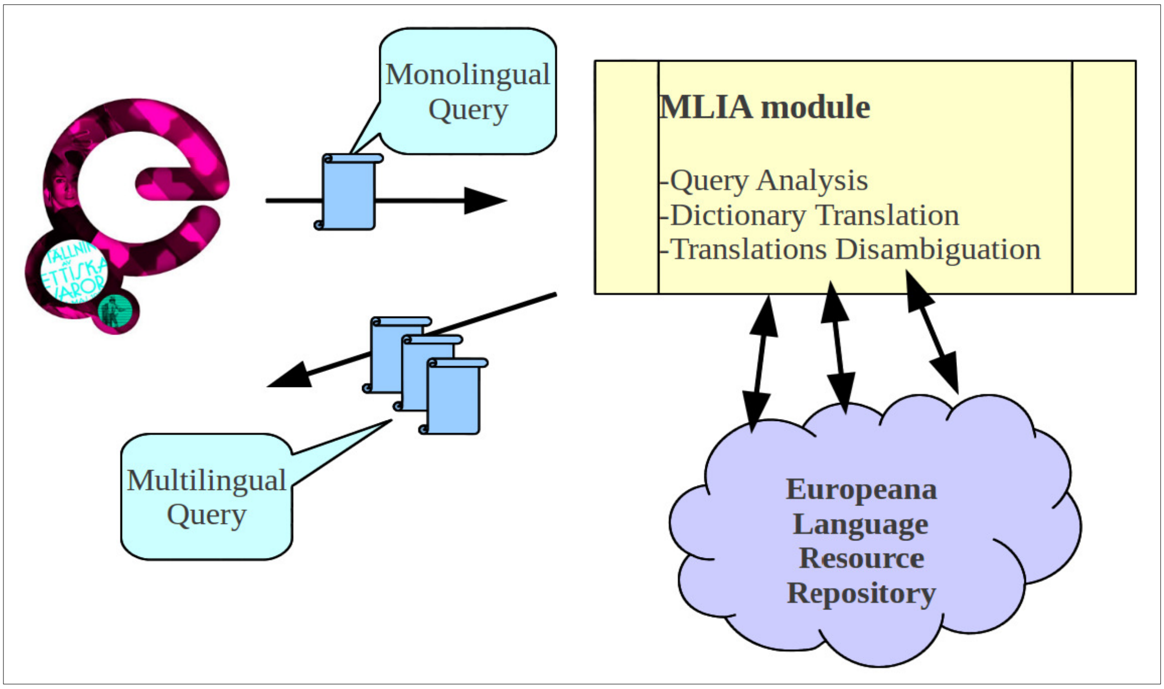
2 MLIA Translation approach

The MLIA module follows a query translation approach based on bilingual dictionaries. The alternative to such an approach is the use of machine translation technology, but dictionary based query translation is preferred in this context for the following reasons:

- MT systems perform in a satisfying way when dealing with syntactically correct systems; however they tend to under-perform in syntactically poor contexts such as web queries.

- In many (probably most) cases translations can be ambiguous. In these cases, most MT systems tend to make a choice, thus blocking the retrieval of potentially interesting digital items; dictionary based techniques, on the contrary, allow for search of (possibly disambiguated) multiple candidate translations;

The translation approach implemented by MLIA module consists in the sequence of 3 different activities (see picture 1), the query analysis, the translation of the query terms by means of bilingual dictionaries and the disambiguation of the translation candidates retrieved by the dictionaries.



Picture 1 – MLIA approach overview

The following subsections detail each one of these 3 sub-tasks.

2.1 Query Analysis

The Query Analysis subtask consists in the identification of the query language (if the information is not explicitly provided by the invoker of the MLIA module), the recognition of Named Entities and the morphological analysis of the query in order to identify the standard form of the terms present in the query in order to enable the access to their translations with the bilingual dictionary lookup.

In the process of query analysis the textual contents of the query (input by the user) is tokenized into different terms (consisting of a single token or more tokens, like multiwords expressions) and each term is enriched with additional pieces of information concerning its POS (Part of Speech), its standard form (i.e. its lemma) and an optional semantic label if the term is recognized as a Named Entity (i.e. PERSON, LOCATION, etc.).

Information about lemma and POS of the input terms are used in order to access the entries within the Dictionary Translation (terms are listed within dictionaries by means of their lemma, the citational form of a word) and in order to filter out translations with a compatible POS (see section 2.2 of the present document on Dictionary Translation for further details)

2.2 Dictionary Translation

The Dictionary Translation subtask consists in retrieving all the translations in a target language available for a given term. The following approach is adopted for term translations:

1. Translate the query terms using their surface form. If the POS feature (or the NE category) of a given source term and the correspondent one of its translation are not compatible, the translation is discarded (i.e. ancora [IT] (ADV) → anchor [EN] (NOUN) is discarded)
2. If no suitable translation is found for a term using its surface form, the MLIA module search for translations using the lemma associated with the term. Unless only a single translation is available in the dictionary, the candidate translations with not compatible POS are discarded.
3. If no suitable translations is found using the surface form or the lemma then if the term has been recognized as a Named Entity is left un-translated otherwise a substring of the original term is used in order to find possible “approximate translations” of the term. Approximate translations are filtered as well using their POS feature.
4. If no translations is found the term is left un-translated.

2.3 Translations Disambiguation

The Translations Disambiguation subtask consists in selecting the most appropriate translations among the candidate ones proposed by the Dictionary Lookup on the basis of the overall context of the query. In order to perform this activity, a semantic component able to associate a semantic vectorial representation to words is needed. The translation disambiguation strategy is employed whenever the input query contains at least two terms.

For such a semantic component we experimented with a corpus-based distributional approach capable of detecting the interrelation between different terms in a corpus; the strategy we adopted is similar to Latent Semantic Analysis (see [1]) although it uses a less expensive computational solution based on the Random Projection algorithm (see [2,3]).

Random Indexing exploits an algebraic model in order to represent the semantics of terms in a Nth dimensional space (a vector of length N); approaches falling into this category, create a “*Terms By Contexts*” matrix where each row represents the degree of memberships of a given term to the different contexts. The Algorithm assigns a random signature to each context (a highly sparse vector of length N, with few, randomly chosen, non zero elements) and then generates the vector space model by performing a statistical analysis of the documents in the domain corpus and by accumulating on terms rows all the signatures of the contexts where terms appear.

According to this approach, if two different terms have a similar meaning they should appear in similar contexts (within the same documents or surrounded by the same words), resulting in close coordinates in the generated semantic space.

In the context of the examples reported in this document, the semantic vectors have been generated taking as corpus Wikipedia articles. After the processing for each word in the corpus we have a vector of floats from 0 to 1 representing its contextual meaning.

In the application scenario of translations disambiguation, the coordinates of the candidate translations available for a given input term (the vector of floats) are compared with the average coordinate of the translations of the other input terms present in the query and ranked by increasing distance. The top elements of this ranked list are then selected as translations.

3 MLIA software interface.

The MLIA module is implemented by the ***QueryTranslation*** class and exposes its query translation / enrichment to the Europena portal by means of a couple of operations.

- *public TranslatedQuery* ***translateQuery****(String query)*
- *public TranslatedQuery* ***translateQuery****(String query, String lang)*

Both operations perform the same task and only differ with respect to input parameters; in the first operation the Language should be autonomously guessed by the system itself while in the second one the Language is explicitly specified with a parameter.

The *TranslatedQuery* object returned by both methods is an utility class used for storing the output of the MLIA module and providing access to the translations by means of the language of the translations or by source query term. In particular the following operations:

- *public String* ***getTranslatedQuery****(Language l):* returns a simple string representation of the translated query.
- *public TreeMap<QueryTerm, Vector<TranslatedTerm>>* ***getQueryTermTranslations****(Language l):* returns a set of translations for each query term present in the input query (individuated in the QueryAnalysis subtask).

The *QueryTerm* and *TranslatedTerm* classes returned by the ***getQueryTermTranslation*** models and provide access to linguistic feature of the source term and the translation candidates (surface form, lemma, part of speech, named entity label, translation confidence,…) A usage example of the MLIA module is reported in Table 1 of subsection 3.2.

These classes are documented in the JavaDoc tree contained in svn repository. The latest documentation can be generated from the source code by going to the directory mlia/ in the repository and running "`mvn javadoc:javadoc`".

3.1 MLIA Configuration

The configuration of the MLIA module is based on standard Spring annotations (see [4]). A specific class acts as configurator and instantiates all the resources needed by the ***QueryTranslator*** class. Different configurator class can be used in order to provide different configuration schemes exploiting different language resources or supporting different languages. Within the context of the project 3 different configuration classes have been implemented:

- TelCLEFtestConfigurationOpenSource creates a ***QueryTranslator*** instance based on Open Source Language resources
- TelCLEFtestConfigurationXEROX creates a ***QueryTranslator*** instance based on XEROX Language resources
- TelCLEFtestConfigurationCELI creates a ***QueryTranslator*** instance based on CELI Language resources

A BaseConfigurator class has been created in the MLIA package in order to group together all the common attributes of configuration classes (input language, target languages, ...) and has been extended by the specific configuration classes. In Subsection 3.2 the usage example shows how to instantiate a ***QueryTranslator*** object using Spring configuration classes.

3.2 Usage Example

Table 1, reported below, presents an usage example of the MLIA module using Portuguese as input and English, French and German as target languages. Two different sentences are used as input:

- Example 1) Emigração irlandesa para a América do Norte
- Example 2) Peregrinação a Santiago de Compostela

and two different methods of the class exposing the results of morphological analysis (TranslatedQuery) are used in the example in order to access the results of the translation activity and print them out to standard output (the output of the morphological analysis of the usage examples is reported in table 2).

In the following example a QueryTranslation object is obtained via Spring framework, then the first example query is translated and the translation in the target language *targetL* is obtained exploiting the method *getTranslatedQuery(Language targetL).* The second example query is then translated and the translations of each term present in the input query is obtained accessing the method *getQueryTermTranslations(Language target)*.

```
/* Retrieve the bean from Container (SPRING is Used to initialize the Query
Translator Bean) */

QueryTranslator qt = (QueryTranslator) context.getBean("queryTranslator");

//get the translations (Example 1)
TranslatedQuery tq = qt.translateQuery("Emigração irlandesa para a América do Norte", l.getId());

//Print on standard output the translations (the whole query translated)
for(Language targetL: TelCLEFtestConfigurationOpenSource.to ){

        String translatedQuery = tq.getTranslatedQuery(targetL);
        System.out.println(translatedQuery+" ["+targetL+"]");
}

//get the translations (Example 2)
tq= qt.translateQuery("Peregrinação a Santiago de Compostela",l.getId());

//Print on standard output the translations (the translations term by term)
for(Language targetL: TelCLEFtestConfigurationOpenSource.to ){

        TreeMap<QueryTerm,Vector<TranslatedTerm>> map=tq.getQueryTermTranslations(targetL);

        for(QueryTerm term: map.keySet()){

                Vector<TranslatedTerm> v = map.get(term);
                System.out.println(term.getSurfaceForm()+" / "+term.getLemma()+"
[POS:"+term.getPos().getUri().getFragment()+"] ("+l.getId()+") ->");

                for(TranslatedTerm tt: v){
                        System.out.println("\t"+tt.getTerm()+"
[POS:"+tt.getPartOfSpeech().getUri().getFragment()+"] ("+targetL.getId()+") ");

                }
        }

}
```

Table 1 – MLIA usage example

The output of the morphological analysis of the usage examples presented in Table 1 (see the System.out.println statements in Table 1) is reported in Table 2.

```
North America emigration Irish [en]

Amérique du Nord émigration Irlandaise [fr]

Nordamerika auswanderung irisch [de]



santiago de compostela / santiago de compostela [POS:UNK] (pt) ->

        Santiago de Compostela [POS:NOUN] (en)

Peregrinação / peregrinação [POS:NOM] (pt ->

        Pilgrimage [POS:NOUN] (en)

        pilgrimage [POS:NOUN] (en)

santiago de compostela / santiago de compostela [POS:UNK] (pt ->

        Santiago de Compostelle [POS:NOUN] (fr)

Peregrinação / peregrinação [POS:NOM] (pt) ->

        pèlerinage [POS:NOUN] (fr)

        Pèlerinage [POS:NOUN] (fr)

santiago de compostela / santiago de compostela [POS:UNK] (pt ->

        Compostela [POS:NOUN] (de)

Peregrinação / peregrinação [POS:NOM] (pt) ->

        Pilgerfahrt [POS:NOUN] (de)

        wallfahrt [POS:NOUN] (de)
```

Table 2 – Output of the morphological analysis

4 Translation Examples

The MLIA module translation capabilities directly depends on the linguistic resources injected in the system and in order to support a given Language proper Language Resources (morphological analyzers, bilingual dictionary lookup) must be present in the LRR repository.

Currently the system supports translations from and to any of the 10 languages supported by EuropeanaConnect: English, French, German, Italian, Spanish, Portuguese, Hungarian, Polish, Netherland, and Swedish.

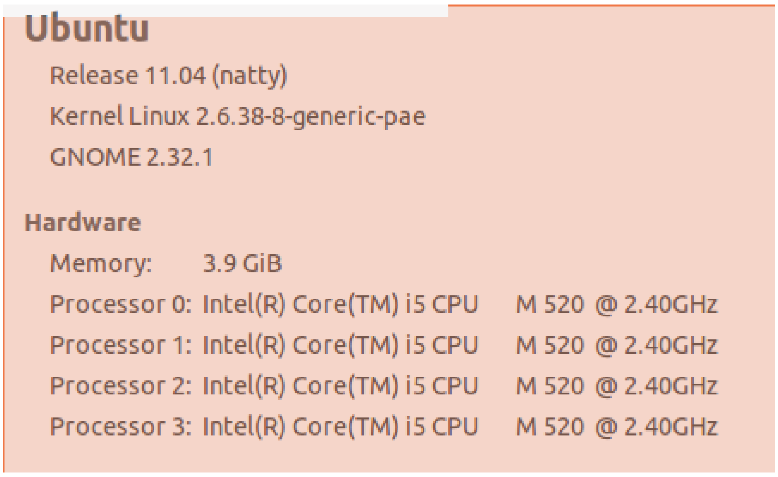
In the followings a set of translation examples generated by the MLIA module is reported. A few queries expressed in each of the 10 Europeana languages are reported along with their translations to English, French, and German. The translations presented are taken from the ongoing tests of the module for the Milestone M2.5.1 [M26]

- Roman Military in Britain (en) -> romain Romain guerrier militaire activiste Britain (fr) - Antiqua Militär militär Britannien (de) - [Elapsed time: 421 milliSec.]
- Modern Japanese Culture (en) -> moderne moderniste Japonais nippon raffinement culture Culture (fr) - modern japanisch Kultur Bildung Zucht (de) - [Elapsed time: 100 milliSec.]
- Famous Jazz Musicians (en) -> musiciens de jazz fameux glorieux célèbre (fr) - jazz-musiker berühmt namhaft berühmte (de) - [Elapsed time: 41 milliSec.]
- Den irländska emigrationen till Nordamerika (sv) -> Irlandaise émigration Amérique du Nord (fr) - Irish emigration North America (en) - irisch auswanderung Nordamerika (de) - [Elapsed time: 508 milliSec.]
- Antika grekiska mynt (sv) -> grec ancien piece jeton pièce de monnaie (fr) - ancient greek coin coins (en) - altgriechisch Geldmünze Geldstück Münze (de) - [Elapsed time: 29 milliSec.]
- Skotsk folkmusik (sv) -> Ecossais musique folk (fr) - Scottish folk music (en) - Schottisch Scotch Folk (de) - [Elapsed time: 35 milliSec.]
- Analyses comparées des systèmes électoraux (fr) -> analysis analytic thinking analysis testzes Electoral systems (en) - Analyse Untersuchung Wahlsystemen (de) - [Elapsed time: 1 sec.]
- Photographies de voitures de collection (fr) -> photography picture taking photograph vintage cars vintage car (en) - Bilder oldtimer (de) - [Elapsed time: 78 milliSec.]
- Recettes de desserts au chocolat (fr) -> returns receipts collector collector Chocolate desserts chocolate desserts (en) - Betriebseinnahmen Einnahmen Erträge Schokoladendesserts schokoladendesserts (de) - [Elapsed time: 74 milliSec.]
- Allergie alimentari (it) -> Allergies alimentaires allergies alimentaires (fr) - Food Allergies food allergies (en) - Nahrungsmittelallergien (de) - [Elapsed time: 1 sec.]
- Afroamericani nella guerra civile in America (it) -> afro-américains guerre civile États-Unis Amérique EU (fr) - african americans civil war Civil war America the Americas (en) - afro-amerikaner Bürgerkrieg Amerika (de) - . [Elapsed time: 532 milliSec.]

- Occupazione giovanile in Europa (it) -> efficace effectif position juvénile Europe europe europe fille d'agénor (fr) - mark post spot youthful young juvenile Europe (en) - Lage Standort Posten jugendlich klein jung Europa (de) - [Elapsed time: 183 milliSec.]
- Emigração irlandesa para a América do Norte (pt) -> Amérique du Nord émigration Irlandaise (fr) - North America emigration Irish (en) - Nordamerika auswanderung irisch (de) - [Elapsed time: 1 sec.]
- Sufrágio feminino nos Estados Unidos (pt) -> unité état suffrage femmes (fr) - united state province America suffrage female (en) - Einheit Staat wahlrecht weiblich (de) - [Elapsed time: 1 sec.]
- Jardinagem para Crianças (pt) -> jardinage horticulture enfants (fr) - gardening children (en) - Gärtnern kinder (de) - [Elapsed time: 63 milliSec.]
- Bombardierung japanischer Städte (de) -> bombardements Japonais japonaises villes (fr) - bombardment shellfire Japanese cities burgs towns (en) - [Elapsed time: 492 milliSec.]
- Irische Emigration nach Nordamerika (de) -> Irlandaise émigration Amérique du Nord (fr) - Irish emigration North America (en) - [Elapsed time: 187 milliSec.]
- Kirchen in Frankreich (de) -> églises République française France (fr) - church metropolitan France (en) - [Elapsed time: 37 milliSec.]
- Skót zene (hu) -> sable sablonneux sableux musique (fr) - Scot music (en) - schottisch Musik (de) - . [Elapsed time: 384 milliSec.]
- Autóipar Európában (hu) -> industrie automobile Industrie de l'automobile Europe (fr) - Auto Industry auto industry europa Europe (en) - Autoindustrie Europa (de) - [Elapsed time: 53 milliSec.]
- Keleti filozófia (hu) -> oriental doctrine philosophe philosophie (fr) - oriental philosophy (en) - oriental orientalisch östlich Philosophie (de) - [Elapsed time: 32 milliSec.]
- podróże po Antarktydzie (pl) -> podróże Antártida (fr) - podróże Antarctica antarctica (en) - podróże Antarktis (de) - [Elapsed time: 588 milliSec.]
- wegetarianizm (pl) -> végétarianisme végétarisme (fr) - vegetarianism (en) - vegetarismus (de) - [Elapsed time: 6 milliSec.]
- produkcja mydła (pl) -> production production de savon (fr) - production manufacture output soap production (en) - Produktion Seifenproduktion (de) - [Elapsed time: 12 milliSec.]
- La inquisición (es) -> inquisition inquisitione (fr) - inquisition (en) - die Inquisition Inquisition Verhör (de) - [Elapsed time: 538 milliSec.]
- Grandes expeciones de cacería África (es) -> grand expeciones cacería afrique (fr) - huge sizeable sizable expeciones pan Africa (en) - groß expeciones cacería afrikanische (de) - [Elapsed time: 287 milliSec.]
- Las esposas de Enrique VIII (es) -> Wives Enrique VIII (fr) - handcuff enrich VIII (en) - Ehefrauen Enrique VIII (de) - [Elapsed time: 138 milliSec.]
- Voedselallergieen (nl) -> allergies alimentaires (fr) - food allergies (en) - Nahrungsmittelallergien (de) - [Elapsed time: 312 milliSec.]

- Zeep maken (nl) -> savon expliquer donner servir (fr) - soap develop (en) - Seife Seifenproduktion machen tun praktizieren (de) - [Elapsed time: 336 milliSec.]
- Grafisch programmeren (nl) -> graphique programmer (fr) - graphic program programme (en) - grafik programmieren (de) - [Elapsed time: 103 milliSec.]

The following tests have been performed on a test environment with the following charachteristics



```
Ubuntu
   Release 11.04 (natty)
   Kernel Linux 2.6.38-8-generic-pae
   GNOME 2.32.1

Hardware
   Memory:       3.9 GiB
   Processor 0:  Intel(R) Core(TM) i5 CPU    M 520  @ 2.40GHz
   Processor 1:  Intel(R) Core(TM) i5 CPU    M 520  @ 2.40GHz
   Processor 2:  Intel(R) Core(TM) i5 CPU    M 520  @ 2.40GHz
   Processor 3:  Intel(R) Core(TM) i5 CPU    M 520  @ 2.40GHz
```

Appendix A. References

1. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 (1990)
2. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: in Knowledge Discovery and Data Mining, ACM Press (2001)
3. Baroni, M., Bisi, S.: Using cooccurrence statistics and the web to discover synonyms in technical language (2004)
4. http://en.wikipedia.org/wiki/Spring_Framework