A Gold Standard for CLIR evaluation in the Organic Agriculture Domain

Alessio Bosca*, Matteo Casu*, Mauro Dragoni**, Nikolaos Marianos***

CELI*, FBK**, AgroKnow***

Turin (Italy)*, Trento (Italy)**, Athens (Greece)

Email: {alessio.bosca, casu}@celi.it, dragoni@fbk.eu, n.marianos@agroknow.gr

Abstract

We present a gold standard for the evaluation of Cross Language Information Retrieval systems in the domain of Organic Agriculture and AgroEcology. The presented resource is free to use for research purposes and it includes a collection of multilingual documents annotated with respect to a domain ontology, the ontology used for annotating the resources, a set of 48 queries in 12 languages and a gold standard with the correct resources for the proposed queries.

The goal of this work consists in contributing to the research community with a resource for evaluating multilingual retrieval algorithms, with particular focus on domain adaptation strategies for “general purpose” multilingual information retrieval systems and on the effective exploitation of semantic annotations. Domain adaptation is in fact an important activity for tuning the retrieval system, reducing the ambiguities and improving the precision of information retrieval. Domain ontologies constitute a diffuse practice for defining the conceptual space of a corpus and mapping resources to specific topics and in our lab we propose as well to investigate and evaluate the impact of this information in enhancing the retrieval of contents.

An initial experiment is described, giving a baseline for further research with the proposed gold standard.

Keywords: Gold Standard, CLIR, Evaluation, Domain Adaptation

1. Introduction

The usage of multilingual resources registered a significant increase in the last decade. Multilinguality is used in several fields of Computer Science and the necessity of managing multilingual information became an important if not critical, task. Different initiatives within institutional and commercial contexts emerged in the last years in order to support multilingual access to information. However the proposed gold standard include:

• A set of 48 parallel queries in 12 languages that express some informational needs in the organic agriculture domain. Each query is identified with a unique ID. Their selection/translation process is described in section 2
• A gold standard that for each query provides a list of correct results, ordered in terms of relevance and that is used to evaluate the results provided by the CLIR system. The gold standard targets the documents expressed in English language within the Organic.Lingua collection and is described in section 3
• A collection of documents that satisfy the informational needs expressed in the queries. Each document in the collection is identified with a unique ID. The collections is constituted by Organic.Lingua multilingual documents and is described in section 4
• The Organic.Lingua ontology used for semantic annotations on the documents, described in section 5

An initial experiment is described in section 6 giving a

1 http://organic-lingua.eu
2 http://aims.fao.org/standards/agrovoc/
3 http://www.clef-initiative.eu
baseline for further research on this corpus. Finally, section 7 concludes the paper.

2. Query Selection Process

The approach adopted for the selection of the queries to include in the evaluation resource consists in the selection of the most popular searches submitted by real users to the portal and includes the following steps:

1. Selection of the queries from search logs: by means of the Google analytics service, a list consisting of the more frequently used queries in the English language was selected from the search logs of the Organic.Edunet Web portal. Domain experts from AgroKnow selected a meaningful subset from the list with the aim of covering as many topics as possible among the ones present in Organic.Lingua documents collections and avoiding similar queries.

2. Manual Translation of the queries: Human domain experts chosen among the Organic.Lingua project partners provided the translations of the queries selected at step 1 from English to the set of target languages (English, French, Italian, Spanish, German, Polish, Portuguese, Hungarian, Turkish, Estonian, Latvian, Greek) in order to obtain a list of parallel queries.

Concerning popular queries selection and multilingual search analytics, the methodology adopted is described by (Stoitis et al, 2011). Table 1 presents an excerpt of the selected queries.

| en                          | fr                          | de                          | es                          | it                          | el                          | tr                          | pl                          | pt                          | hu                          | lv                          | et                          |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| organic certification      | certification biologique    | biologisches Zertifikat     | certificación ecológica     | certificazione biologica    | βιολογική πιστοποίηση         | organik sertifikasyon       | certyfikacja produkcji organicznej | certificação orgânica       | ökológiai minősítés           | biologiská sertifikácia    | mahepõllun duse sertifikaat |
| alternative crops          | cultures alternatives       | alternative                    | cultivos alternativos       | colheitas alternativas      | αναλλακτικές καλλιέργειες     | alternatif ürünler          | rośliny uprawne               | alternativas                | alternatívnyek               | alternatīvās kultūras      | alternativne saak          |
| animal manure              | fumier                      | Stalldung                    | estiercol                   | letame                      | ζωική κοπριά                   | hayvan gübresi              | nawóz zwierzęcy              | adubo                        | trägya                      | kūtsmēšļi                   | loomasõnnik                 |
| food safety                | sécurité alimentaire        | Lebensmittelbioverarbeitung  | seguridad                   | sicurezza alimentare        | ὁμορράζως τροφίμων             | gda güvenliği               | bezpieczenś wo żywności      | segurança alimentar         | ēlelmiszerbi ztonsāg         | nekaifiga pārtika          | toida ohutus                |

Table 1: Examples of selected queries

3. Gold Standard for Selected Queries

The evaluation resource contains a gold standard that associates to each query a list of documents considered to be relevant for the informational needs of the query. The gold standard is compiled for the English queries and targets only the documents expressed in the English language within the collection (as the documents coverage for the other languages supported in the evaluation was not wide enough for building a gold standard). The CLIR task proposed with our evaluation resource follows the model-independent approach and treats translation and retrieval as two separate processes, therefore it requires the translation of queries from a language XX to English followed by a monolingual retrieval phase.

The documents are referenced in the gold standard through the univoque value of the General Identifier metadata field.

4. Organic.Lingua Learning Objects

Metadata

Organic Lingua collection consists of about 13,000 records with high quality metadata (human edited semantic annotations and metadata translations). The records in the collection consist of metadata describing on-line educational contents for the awareness and education of European youth about organic agriculture and agroecology expressed in different media and formats.
(web pages, pictures, pdf documents, videos) and are specified according to the IEEE LOM (Learning Object Metadata) standard. The multilingual collection of records is expressed in different languages (English, French, German, Danish, Hungarian, Norwegian, Romanian, Russian, Greek, Spanish, Estonian, Swedish, Portuguese, Arabic, Hindi) and the coverage of documents for the different languages is not regular and one language can be frequent while others are quite sparse. The collection is provided as XML encoded files.

5. Organic Lingua Ontology

The Organic.Lingua ontology was created by an expert committee on OA&AE topics in the framework of the Organic.Edunet project. Experts studied existing taxonomies and categorizations – such as the BIOAGRO OA taxonomy, FAO’s AGRIS/CARIS subject categories and AGROVOC – and used them as a basis to elaborate a classified list of terms that later evolved into the current ontology following a process described in detail in (Holm et al., 2010). The Organic.Edunet ontology was designed to help users of the Organic.Edunet portal to have more efficient searches (called semantic searches). It is therefore application-specific and dependent. According to the classification by (Guarino, 1998), it is an application ontology, as it describes “concepts depending both on a particular domain and task”.

The current ontology is multilingual (15 languages, namely English, Spanish, German, Greek, Romanian, Hungarian, Estonian, Russian, Norwegian-Bokmål, Bulgarian, Czech, Turkish, French, Hindi, Slovenian and Arabic) and it is implemented in OWL 2. The evaluation resource includes as well a SKOS representation of the same ontology besides the OWL 2 one. Table 2 summarizes some basic data about the ontology while Table 3 presents the coverage of concepts labels and descriptions with respect to the supported languages.

| Language | Labels | Descriptions |
|----------|--------|--------------|
| Arabic   | 281/375| 0/375        |
| Bulgarian| 297/375| 0/375        |
| Dutch    | 281/375| 0/375        |
| English  | 372/375| 88/375       |
| Estonian | 373/375| 368/375      |
| French   | 370/375| 205/375      |
| German   | 281/375| 0/375        |
| Greek    | 373/375| 365/375      |
| Hindi    | 281/375| 0/375        |
| Hungarian| 281/375| 0/375        |
| Italian  | 371/375| 366/375      |
| Latvian  | 308/375| 304/375      |
| Norwegian| 281/375| 0/375        |
| Romanian | 281/375| 0/375        |
| Russian  | 373/375| 372/375      |
| Slovakian| 281/375| 0/375        |
| Spanish  | 369/375| 367/375      |
| Turkish  | 374/375| 368/375      |

Table 3: Organic.Lingua ontology - language coverage

It is important to say that most learning objects are classified by an average of 4 ontology terms and not with a single classification value. Such richness in semantic annotations allows for investigating different strategies in exploiting the ontological resource. Illustration 1 presents the distribution of terms in the ontology according to the number of learning objects they classify.

Illustration 1: Concepts distribution over annotated resources

6. Initial Experiment

In this section we present the experimental results obtained by testing Organic.Lingua CLIR module with the proposed evaluation resource. The queries in the gold standard have been translated from the available

---

4 http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
5 http://organic-edunet.eu/
languages to English by using the translation module of Organic.Lingua CLIR system and the obtained translations have been used to perform the retrieval from the Organic.Lingua documents collections. The CLIR system has been evaluated by adopting two different configurations:

1. Base configuration: each query is translated in English by using the CLIR system and it is performed on the textual field of the indexed documents.

2. Configuration with Semantic Expansion: this setting exploits the multilingual ontology labels used for enriching the representation of each document. Each query is translated in English and it is performed on both the textual and the annotation fields of the indexed documents.

Besides the cross-language experiments, an English monolingual run was performed in order to provide a baseline for performance comparison. Table 4 and Table 5 present the results of the performed evaluation, grouped by configuration type.

| Table 4: Base configuration results |
|-------------------------------------|
| **Lang** | **Map** | **Precision @5** | **Precision @10** | **Precision @20** | **Recall** | **R-Precision** |
| en | 0.7261 | 0.7917 | 0.6896 | 0.5865 | 0.9635 | 0.6897 |
| el | 0.3731 | 0.3833 | 0.3479 | 0.3104 | 0.8253 | 0.3756 |
| lv | 0.3348 | 0.325 | 0.3187 | 0.2948 | 0.703 | 0.3483 |
| pl | 0.2559 | 0.3 | 0.2708 | 0.2552 | 0.678 | 0.2671 |
| it | 0.4175 | 0.4208 | 0.3729 | 0.3458 | 0.813 | 0.4223 |
| fr | 0.3557 | 0.4042 | 0.3568 | 0.3193 | 0.7915 | 0.3545 |
| tr | 0.3478 | 0.3917 | 0.3646 | 0.3482 | 0.8134 | 0.3486 |
| hu | 0.2406 | 0.2667 | 0.2708 | 0.251 | 0.6898 | 0.2385 |
| et | 0.3263 | 0.3667 | 0.3438 | 0.3281 | 0.6234 | 0.3596 |
| de | 0.2362 | 0.2458 | 0.1979 | 0.1906 | 0.6436 | 0.2549 |
| es | 0.358 | 0.4042 | 0.3521 | 0.3042 | 0.8356 | 0.3498 |
| pt | 0.5048 | 0.5708 | 0.4896 | 0.425 | 0.904 | 0.4807 |

Table 5: Semantic expansion configuration results

| **Lang** | **Map** | **Precision @5** | **Precision @10** | **Precision @20** | **Recall** | **R-Precision** |
|------------|-----------|------------------|------------------|------------------|------------|----------------|
| en | 0.7351 | 0.7667 | 0.6875 | 0.5906 | 0.9803 | 0.6826 |
| el | 0.37 | 0.4292 | 0.3896 | 0.3448 | 0.8412 | 0.343 |
| lv | 0.3429 | 0.3917 | 0.35 | 0.3198 | 0.7059 | 0.3451 |
| pl | 0.2698 | 0.3417 | 0.3062 | 0.2708 | 0.7084 | 0.2692 |
| it | 0.3972 | 0.4458 | 0.3792 | 0.3323 | 0.8266 | 0.3675 |
| fr | 0.3587 | 0.4167 | 0.4027 | 0.3402 | 0.7961 | 0.3588 |
| tr | 0.3331 | 0.425 | 0.375 | 0.3398 | 0.8297 | 0.3412 |
| hu | 0.2167 | 0.2917 | 0.2792 | 0.2344 | 0.7152 | 0.2184 |
| et | 0.3177 | 0.425 | 0.375 | 0.3398 | 0.6363 | 0.3394 |
| de | 0.2217 | 0.2792 | 0.25 | 0.2406 | 0.6409 | 0.2427 |
| es | 0.3708 | 0.4458 | 0.4167 | 0.3573 | 0.8518 | 0.3591 |
| pt | 0.4633 | 0.55 | 0.4729 | 0.4219 | 0.9099 | 0.4504 |

7. Conclusions

In this short paper we presented a novel evaluation resource for CLIR system in a specific domain. The described resource is hosted on Organic.Lingua project portal and its use is free for evaluation, research and teaching purposes.

8. Acknowledgements

This work has been supported and founded by Organic.Lingua EU project.

9. References

M. Braschler, M., Peters, C. (2003). CLEF 2002 Methodology and Metrics. Advances in Cross-Language Information Retrieval, Lecture Notes in Computer Science, Volume 2785/2003, pp. 512-528.

Stoitis, G., Manouselis, N., Sanchez-Alonso, S. (to be published). Data set requirements for multilingual learning analytics. Int. J. Technology Enhanced Learning.

Holm, A.S., Sanchez-Alonso, S., Sicilia, M.A., Lieblein, G. (2010). Revised version of Organic Agriculture and Agroecology Domain Model Representation, deliverable D2.2.3c. Organic.Edunet project.

Guarino, N. (1998). Formal ontology and information systems. Formal Ontology in Information Systems. IOS Press, pp. 3-15.

---

6 http://www.organic-lingua.eu/deliverables/OrganicLingua_CLIR_Evaluation.zip