Using Semantic and Domain-Based Information in CLIR Systems

Alessio Bosca[1], Matteo Casu[1], Mauro Dragoni[2], and Chiara Di Francescomarino[2]

[1] Celi s.r.l., via S.Quintino 31, 10131, Torino, Italy
[2] FBK–IRST, Trento, Italy
{alessio.bosca,casu}@celi.it, {dragoni,dfmchiara}@fbk.eu

Abstract. Cross-Language Information Retrieval (CLIR) systems extend classic information retrieval mechanisms for allowing users to query across languages, i.e., to retrieve documents written in languages different from the language used for query formulation. In this paper, we present a CLIR system exploiting multilingual ontologies for enriching documents representation with multilingual semantic information during the indexing phase and for mapping query fragments to concepts during the retrieval phase. This system has been applied on a domain-specific document collection and the contribution of the ontologies to the CLIR system has been evaluated in conjunction with the use of both Microsoft Bing and Google Translate translation services. Results demonstrate that the use of domain-specific resources leads to a significant improvement of CLIR system performance.

Keywords: #eswc2014Bosca.

1 Introduction

Cross-Language Information Retrieval (CLIR) deals with the problem of finding documents written in a language different from the one used for query formulation. If attempts to model multilinguality in information retrieval date back to the early Seventies [1], a renewed interest was brought to the field by the rise of the Web in the mid-Nineties, when pages written in different languages started to become suddenly available to geographically distributed users of the Web. International organizations, governments of multi-lingual countries, to name the most important ones, have been traditional users of CLIR systems. In the last decade, however, with the growth in the number of Web users, the need of facing the problem of the language barriers for exchanging information has notably increased and the need for CLIR systems in everyday life has become more and more clear (the recent book by J.-Y. Nie [2] exposes in detail the need for cross-language and multilingual IR).

There are several ways to cross the language barriers in CLIR systems. All of them, however, have to deal with the problem of the language mismatch between the queries and, at least, part of the document content. We can group the possible CLIR scenarios into the following three main settings:

1. the document collection is monolingual, but users can formulate queries in more than one language.

V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 240–254, 2014.

2. the document collection contains documents in multiple languages and users can query the entire collection in one or more languages.
3. the document collection contains documents with mixed-language content and users can query the entire collection in one or more languages.

In this paper, we present an approach facing the third scenario. The proposed CLIR system manages a collection of documents containing multilingual information as well as user queries that may be performed in any language supported by the system. The discussed approach uses domain-specific ontologies for increasing the effectiveness of already-available machine translation services (like Microsoft Bing[1] and Google Translate[2]) by expanding the queries with concepts coming from the ontologies.

The originality of the implemented system consists of the combination of two crucial aspects: (i) domain-specific multilingual ontologies are used for performing query expansion operations; and (ii) these ontologies are exploited also for enriching the representation of documents within the index. This way, the gained benefit expected by the proposed approach is twofold: an improvement of the effectiveness of the ranks produced by the CLIR systems; and the evidence that multilingual ontologies help to have an accurate enrichment of document representation.

The remainder of the paper is structured as follows. Section 2 presents an overview of the works carried out in the field of CLIR systems. Section 3 describes the implemented system and the algorithms used for indexing documents and for computing the IR relevance score. In Section 4 we show how the evaluation has been set up; while, in Section 5 we discuss the obtained results. Finally, Section 6 concludes.

2 Related Work

The problem of multilingual text retrieval has a long history. First experiments on multilingual text retrieval systems, based on the use of bilingual thesaurus, were performed by Salton [1]. Although the proposed approaches are no more feasible in modern systems, their underlying rationale is the basis of modern approaches that use Machine-Readable Dictionaries (MRD). Such approaches use controlled vocabularies for translating terms at query or indexing time. Examples of these approaches are presented in [3] and [4] where frequency statistics are used for selecting the translation of a term; contrariwise, in [5] and [6] more sophisticated techniques exploiting term co-occurrence statistics are described.

MRD-based approaches demonstrated to be effective for addressing the CLIR problem; however, when CLIR systems are applied to specific domains, they suffer of the "Out-Of-Vocabulary" (OOV) issue [7]. OOV problem consists of having a dictionary that is not able to completely cover all terms of a language or, more generally, of a domain. Several studies recognized that the problem of translating OOV has a significant impact on the performance of CLIR systems [8,9]. This problem has been addressed in two different ways in the literature. A first group of approaches [10,11] relies on augmenting the translation lexicon by mining comparable corpora. A second set of

[1] `http://www.bing.com/translator`
[2] `translate.google.com`

approaches, instead, employs Machine Transliteration systems to transliterate proper nouns. Discussions about this strategy are presented in [12] and [13].

Contrarily to the OOV approaches, CLIR domain-based approaches aim at providing systems that, by adapting themselves to a particular domain, are able to obtain higher effectiveness values due to their higher coverage of domain specific terms. For example, in [14], the authors present an approach in which they exploit domains, coming from Web directories, for providing better translations of queries. In [15] a cross-language medical information retrieval system has been implemented by exploiting for translations, a thesaurus enriched with medical information. In both works, the results demonstrated that the idea of using domain specific resources for CLIR is promising.

Recently, approaches exploiting the use of semantics have been explored. Such approaches enrich the document representation by injecting in the index of each document, a set of concepts coming from thesauri and/or ontologies in order to facilitate the cross-language retrieval of the document itself [16].

This work falls in both the last two streams of works, borrowing from the former the advantages deriving from the usage of domain-specific terms in the query translation and from the latter the capability to exploit semantic knowledge for retrieving information.

Other specific works on CLIR within the multilingual semantic web may be found in [17] and [18], while a complete overview of the ongoing research on CLIR is available at the Cross-Language Evaluation Forum (CLEF[3]), one of the major references concerning the evaluation of multilingual information access systems.

3 Approach and Implemented System

In this section, we describe the approach we have adopted for addressing the CLIR problem. Since the main goal of the presented work consists of exploring the impact of domain-specific semantic resources on the effectiveness of CLIR systems, in our investigations we will focus on the strategies for matching textual inputs to ontological concepts (applied to both the query and the documents in the target collection) rather than on the translation of the textual query.

The system described in this work has been developed in the context of the Organic.Lingua[4] EU-funded project that aims at providing automated multilingual services and tools facilitating the discovery, retrieval, exploitation and extension of digital educational contents related to the domain of Organic Agriculture and AgroEcology. The proposed approach supports two terminology resources: the multilingual ontology from the Organic.Edunet portal[5] (specifically developed in the context of the project for annotating documents) and a more generic resource, but domain-specific, namely Agrovoc that is a multilingual theasurus from FAO[6]. Both resources are expressed with SKOS format.

[3] http://www.clef-initiative.eu
[4] http://www.organic-lingua.eu
[5] http://organic-edunet.eu/
[6] http://aims.fao.org/standards/agrovoc/about

The system presented in this paper follows the Model-independent approach and treats translation and retrieval as two separate processes. The queries are first translated into the document language and monolingual IR models are then directly applied. A typical and also broadly used approach of this type is the machine translation (MT) approach (e.g.[19]) which employs MT systems to translate queries or documents before applying the monolingual retrieval process. In our implementation we followed such an approach for query translation and exploited Google Translate[7] and Bing Translator[8] as MT services.

In the subsections below, we describe how the proposed system performs the document indexing and their retrieval.

3.1 From Query Terms to Concepts

The component for matching a textual input with elements from domain terminologies is based on the Search Engines technology and exploits its built-in textual search capabilities. In our implementation, we exploited the open source Lucene search engine[9] and created a search index for each of the supported languages, containing the textual labels of the terminology elements (both SKOS preferred labels and alternative ones) along with their URI. The terms labels are indexed in their original form as well as in their stemmed form by means of the default stemming resources available in the Lucene framework.

In order to find the terminological entries within a textual input expressed in a given language a two steps procedure is applied:

- At first, the text is used as a query and is searched over the index in order to find a list of all the terminology elements containing a textual fragment present in the text.
- As a second step, in order to retain only the domain terms with a complete match (no partial matches) and locate them in the text, a new search index is built in memory, containing a single document: the original textual input. Then the candidate terminology elements found in the first step are used as queries over the in-memory index and a Highlighter component of the Search Engine is exploited to locate them in the text. A longest match criterion is used when the found terminology elements refer to overlapping spans of text.

This procedure is applied at indexing time in order to find references to ontological concepts within the textual fields of the documents and at query time in order to locate domain concepts in the query submitted by users. In the retrieval phase, the conceptual references found in the query are matched against the concepts annotated in the indexed documents.

3.2 Indexing

In order to compute the document index, each field with textual contents is extracted from the documents. Stop-word removal and stemming algorithms suited for each spe-

[7] http://translate.google.com
[8] http://www.bing.com/translator
[9] https://lucene.apache.org

cific language are applied to the fields before indexing them. The procedure for textual match described in the previous section allows for the enrichment of documents with annotations referencing ontology concepts.

Besides the annotations computed automatically, the original collection of documents exploited in our experimental evaluation already includes manual annotation with respect to the Organic.Lingua domain ontology (described in Section 4).

Moreover, in order to store into the index the information related to the context of each conceptual annotation, each concept used for annotating the document is expanded by considering its ontological parents and by indexing them according to a decreasing weight that depends on their semantic distance from the concept [20]. Therefore, the final representation of each document in the index is given by textual fields (exploited for the textual search) and annotations fields (exploited for the conceptual search). All fields are indexed by using the Lucene variation of the TF-IDF model.

Table 1 presents a statistic of the manual and automatic semantic annotations created at indexing time.

Table 1. Statistics of Manual and Automatic Conceptual Annotations performed at indexing time

Domain Ontology	Number of Concepts	Manual Annotations	Automatic Annotations
Agrovoc	32061	0	133596 annotations about 5834 distinct concepts
Organic.Lingua	291	27871 about 264 distinct concepts	16434 annotations about 208 distinct concepts

3.3 Retrieval

The proposed CLIR system provides two different components for transforming the queries formulated by users into the final ones performed on the index. These components interact, respectively, with the MT services and with the domain-specific ontology deployed on the CLIR system. At query time, the CLIR system may perform the construction of three types of queries, starting from the ones formulated by users, based on the system configuration:

1. Only Translations: query terms are translated into the reference language used for retrieving documents.
2. Only Semantic: for each query term, the CLIR system looks for a match into the ontology. If a match is found, the concept is put into the semantic transformation of the original query, together with its parent concepts extracted from the ontology; otherwise, the term is discarded from the final query. This way, the final query will be composed only by the list of the terms for which an ontological match is found, plus the list of concepts representing their contexts.
3. Translation + Semantic: the final query is the combination of the two approaches described above. Therefore, given a list of query terms, they are both translated in the reference language, and matched with ontology concepts. The result is a query composed of three parts: the translation of the original query, the set of concepts matching the terms contained in the original query, and their semantic context.

4 Experiments Setup

In this section, we describe the concrete exploitation of multilingual ontologies in a cross-language resource retrieval use-case in the context of the Organic.Lingua project. The evaluation of the proposed approach has been inspired by the activities of the CLEF, one of the major references concerning the evaluation of multilingual information access systems. Based on this methodology, the resources used for such an evaluation include[10]:

1. A set of queries that express information needs in a given language identified with a unique ID. The approach adopted for selecting the queries consisted of choosing the most popular searches performed by real users on the Organic.Lingua portal filtered by domain experts. This way, we are able to cover as many topics as possible, while avoiding similar queries. The number of queries used for these experiments is 48. Each query has been originally provided in the English language and it has been manually translated in the set of the other languages and verified by both Domain and Language experts.
2. A collection of documents that satisfies the information needs expressed in the queries. In the Organic.Lingua test environment this corpus is composed of a multilingual collection of about 13000 documents.
3. A gold standard that, for each query, provides the list of the relevant documents used to evaluate the results provided by the CLIR system. In the provided evaluation, the gold standard was manually created by the domain experts. It contains only results that are related to queries expressed or translated in English and that have at least one field (either a textual or an annotation one) in English.

4.1 Evaluation Metrics

For evaluating the effectiveness of the CLIR system, different standard metrics have been adopted. Besides the well-known Precision and Recall measure, other metrics are widely used in the IR community. By keeping as reference the CLEF evaluation campaigns, the metrics used in recent years include R-Precision, Precision@X (representing the Precision obtained after X retrieved documents, i.e. Prec@10 is the precision after 10 docs) and the Mean Average Precision (MAP). Since the evaluation of the Organic.Lingua CLIR system is based on the methodology introduced by CLEF [21,22], the same metrics will be used for evaluating the described system.

5 Evaluation and Discussion

The set of topics considered in the experiment is composed of queries in 8 different languages: French, Italian, Spanish, German, Polish, Portuguese, Hungarian, Turkish. The queries have been translated in English by using the external machine translation

[10] All the evaluation resources are freely available online for reproducing the experiments: `http://www.organic-lingua.eu/deliverables/OrganicLingua_CLIR_Evaluation.zip`

services connected with the CLIR system and then, they have been enriched with concepts coming from the ontologies that match query terms. Finally, queries are performed on the Organic.Lingua document collections. The CLIR system has been evaluated by adopting three different configurations and the results have been compared with the gold standard, according to the metrics described above.

1. *Query Translation configuration*: each query is translated in English by using the Microsoft Bing Translator or the Google Translate service, and the retrieval is performed on the textual fields (i.e., title, abstract and content; while, fields containing media data that are present in some documents have not been considered in our work) of the indexed documents. This configuration permitted to define the baseline of our experiments (Table 2).
2. *Semantic Expansion by exploiting the domain ontology*: this configuration combines the previous ones with the term match approach described in Section 3. Each query is translated and its terms are mapped to the Domain Ontology (Sections 5.1, 5.2, and 5.3). Retrieval is performed both on the textual fields and on the ontological annotation fields (manual, automatic, or both depending on the configuration) of the indexed documents.
3. *Ontology Matching Only configuration*: each query term is mapped to one or more concepts of the Domain Ontology by using the approach described in Section 3 and only queries containing at least one match to the Domain Ontology are performed on the index. Retrieval is performed both on the textual fields and on the ontological annotation fields of the indexed documents (Section 5.4).

Moreover, for the second and third configurations, different variants, described in more detail in the following subsections, have been applied.

Tables from 2 to 12 report the results of the performed evaluation split in different subsections based on the configuration type. Table 2 reports only data referring to the baseline that we adopted for comparing the proposed approach. The columns of each table show the Mean Average Precision, the Precisions at 5, 10, 20, and 30, the Average Recall, the Average R-Precision, and the number of queries that have been performed.

In the following subsections, we will present the results obtained with the different configurations adopter for evaluating the proposed CLIR system.

5.1 Semantic Expansion with Automatic Annotations Only

In this experiment, queries are performed only on document fields containing automatic annotations. In particular, we have explored three variants; queries have been expanded by exploiting (i) only the Agrovoc ontology (Table 3), (ii) only the Organic.Lingua ontology (Table 4) ,or (iii) both ontologies (Table 5).

The results obtained by performing the annotation of documents and the expansion of queries by using only automatic annotations highlight that the use of the ontologies leads to an improvement of the system effectiveness. A first important aspect to observe, is that the sole use of the Agrovoc ontology gives a higher contribution with respect to the sole use of the Organic.Lingua one as it may be inferred from the δ values. The reason is given by the highest coverage of the Agrovoc ontology with respect to the Organic.Lingua one.

Table 2. Baseline results obtained by translating the queries using public available machine translations services like Microsoft Bing and Google Translate without using semantic expansion techniques

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.681	0.681	0.742	0.742	0.644	0.644	0.553	0.553	0.970	0.970	0.653	0.653	48
el	0.519	0.555	0.604	0.592	0.517	0.523	0.431	0.463	0.923	0.933	0.514	0.551	48
lv	0.581	0.551	0.629	0.613	0.560	0.542	0.465	0.445	0.960	0.958	0.563	0.537	48
pl	0.540	0.540	0.617	0.617	0.550	0.550	0.468	0.468	0.921	0.921	0.533	0.533	48
it	0.605	0.613	0.675	0.675	0.579	0.594	0.481	0.509	0.942	0.903	0.599	0.597	48
fr	0.513	0.475	0.567	0.517	0.513	0.477	0.441	0.398	0.917	0.863	0.494	0.470	48
tr	0.477	0.456	0.550	0.508	0.494	0.471	0.447	0.413	0.898	0.885	0.466	0.445	48
hu	0.482	0.531	0.563	0.583	0.521	0.542	0.457	0.465	0.895	0.910	0.475	0.515	48
et	0.462	0.495	0.546	0.554	0.471	0.469	0.397	0.407	0.871	0.866	0.451	0.490	48
de	0.564	0.527	0.613	0.588	0.540	0.513	0.449	0.438	0.904	0.886	0.538	0.510	48
es	0.598	0.623	0.671	0.688	0.596	0.598	0.508	0.514	0.936	0.959	0.591	0.613	48
pt	0.616	0.614	0.704	0.671	0.608	0.579	0.496	0.483	0.951	0.942	0.607	0.605	48
AVG.	0.553	0.555	0.623	0.612	0.549	0.542	0.466	0.463	0.924	0.916	0.540	0.543	

Table 3. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from Agrovoc ontology

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.692	0.692	0.746	0.746	0.671	0.671	0.566	0.566	0.972	0.972	0.663	0.663	48
el	0.535	0.568	0.617	0.608	0.540	0.540	0.457	0.484	0.933	0.957	0.524	0.556	48
lv	0.593	0.572	0.638	0.633	0.585	0.567	0.482	0.462	0.965	0.965	0.578	0.563	48
pl	0.561	0.561	0.654	0.654	0.588	0.588	0.488	0.488	0.941	0.941	0.545	0.545	48
it	0.627	0.623	0.708	0.688	0.608	0.613	0.497	0.519	0.944	0.938	0.615	0.605	48
fr	0.532	0.510	0.583	0.533	0.554	0.492	0.463	0.419	0.946	0.929	0.508	0.497	48
tr	0.491	0.478	0.563	0.533	0.519	0.502	0.457	0.433	0.914	0.923	0.481	0.470	48
hu	0.500	0.552	0.588	0.613	0.552	0.575	0.475	0.484	0.923	0.933	0.494	0.533	48
et	0.494	0.517	0.579	0.588	0.513	0.504	0.429	0.437	0.924	0.921	0.481	0.521	48
de	0.582	0.549	0.642	0.613	0.563	0.548	0.463	0.456	0.924	0.935	0.569	0.546	48
es	0.610	0.634	0.679	0.700	0.619	0.617	0.522	0.522	0.951	0.965	0.595	0.613	48
pt	0.631	0.627	0.704	0.671	0.629	0.598	0.510	0.495	0.960	0.964	0.623	0.612	48
AVG.	0.571	0.573	0.642	0.632	0.578	0.568	0.484	0.480	0.941	0.945	0.556	0.560	48
δ w.r.t. Baseline (%)	3.161	3.316	2.952	3.173	5.283	4.808	3.857	3.772	1.899	3.156	2.979	3.142	

Since manual annotations have not been performed on the Agrovoc ontology, the results shown in Table 3 are the same for all running configuration (except the one using only ontological concepts for performing queries). Therefore, they will not be reported in the next two subsections.

5.2 Semantic Expansion with Automatic and Manual Annotations (Same Weights)

In this experiment, we have performed queries on both the fields containing automatic and those containing manual annotations. In this case, we have explored only two variants; queries are expanded by exploiting (i) only the Organic.Lingua ontology (Table 6) or (ii) both ontologies (Table 7).

Indeed, the evaluation adopting only the Agrovoc ontology is not available because this ontology has not been exploited for annotating documents manually.

Table 4. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from Organic.Lingua ontology.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.683	0.683	0.733	0.733	0.652	0.652	0.559	0.559	0.970	0.970	0.653	0.653	48
el	0.527	0.560	0.608	0.588	0.525	0.535	0.442	0.468	0.924	0.933	0.511	0.550	48
lv	0.594	0.569	0.638	0.625	0.588	0.567	0.482	0.462	0.965	0.961	0.574	0.556	48
pl	0.554	0.554	0.633	0.633	0.565	0.565	0.491	0.491	0.931	0.931	0.545	0.545	48
it	0.611	0.617	0.683	0.671	0.596	0.608	0.497	0.518	0.944	0.908	0.602	0.599	48
fr	0.539	0.504	0.596	0.529	0.546	0.477	0.466	0.407	0.930	0.907	0.526	0.503	48
tr	0.489	0.480	0.563	0.546	0.504	0.500	0.453	0.442	0.905	0.909	0.487	0.470	48
hu	0.500	0.540	0.600	0.588	0.546	0.552	0.467	0.473	0.889	0.908	0.488	0.519	48
et	0.487	0.509	0.579	0.579	0.500	0.485	0.428	0.424	0.883	0.871	0.476	0.504	48
de	0.576	0.542	0.638	0.608	0.554	0.533	0.460	0.450	0.910	0.893	0.556	0.524	48
es	0.607	0.632	0.675	0.692	0.604	0.608	0.517	0.524	0.938	0.962	0.594	0.619	48
pt	0.625	0.622	0.700	0.671	0.621	0.585	0.509	0.492	0.954	0.954	0.613	0.609	48
AVG.	0.566	0.568	0.637	0.622	0.567	0.556	0.481	0.476	0.929	0.926	0.552	0.554	48
δ w.r.t. Baseline (%)	2.341	2.288	2.225	1.586	3.163	2.594	3.186	2.780	0.522	1.024	2.223	2.031	

Table 5. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from both Agrovoc and Organic.Lingua ontologies.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.691	0.691	0.746	0.746	0.669	0.669	0.567	0.567	0.972	0.972	0.661	0.661	48
el	0.535	0.567	0.617	0.608	0.538	0.538	0.458	0.485	0.933	0.957	0.522	0.554	48
lv	0.592	0.572	0.638	0.633	0.583	0.565	0.482	0.462	0.965	0.965	0.576	0.562	48
pl	0.560	0.560	0.654	0.654	0.585	0.585	0.488	0.488	0.941	0.941	0.543	0.543	48
it	0.626	0.622	0.708	0.688	0.606	0.610	0.497	0.519	0.944	0.938	0.613	0.604	48
fr	0.532	0.510	0.583	0.533	0.554	0.492	0.463	0.419	0.946	0.929	0.508	0.497	48
tr	0.490	0.477	0.563	0.533	0.517	0.500	0.457	0.433	0.914	0.923	0.480	0.469	48
hu	0.497	0.550	0.583	0.608	0.546	0.569	0.474	0.483	0.923	0.933	0.488	0.527	48
et	0.494	0.517	0.579	0.588	0.508	0.500	0.431	0.438	0.924	0.921	0.478	0.518	48
de	0.582	0.549	0.638	0.608	0.563	0.548	0.463	0.456	0.924	0.935	0.569	0.546	48
es	0.610	0.633	0.679	0.700	0.617	0.615	0.522	0.522	0.951	0.965	0.594	0.612	48
pt	0.630	0.626	0.704	0.671	0.627	0.596	0.510	0.495	0.960	0.964	0.622	0.611	48
AVG.	0.570	0.573	0.641	0.631	0.576	0.565	0.484	0.480	0.941	0.945	0.555	0.559	48
δ w.r.t. Baseline (%)	3.055	3.210	2.840	3.059	4.871	4.391	3.914	3.808	1.899	3.156	2.669	2.829	

The introduction of manual annotations done with the concepts defined in the Organic.Lingua ontology boosted the effectiveness of CLIR system. Indeed, if we compare the δ values obtained by running this configuration, with respect to the ones obtained with the previous configuration, we observe that the gain registered with the use of the sole Organic.Lingua ontology significantly improved. This positive improvement affects also the combined use of the two ontologies for both annotating documents and querying the repository. As for the previous configuration, the highest gain with respect to the baseline is observed for the Prec@10 values, but, in general, there are significant improvements also for the other Prec@X values.

5.3 Semantic Expansion with Automatic and Manual Annotations (Different Weights)

Also in this experiment, we have performed queries on both the fields containing automatic and those containing manual annotations and we have explored two variants

Table 6. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from Organic.Lingua ontology.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.676	0.676	0.742	0.742	0.650	0.650	0.562	0.562	0.972	0.972	0.646	0.646	48
el	0.545	0.565	0.642	0.608	0.550	0.544	0.463	0.472	0.928	0.935	0.528	0.552	48
lv	0.601	0.571	0.654	0.633	0.594	0.573	0.496	0.472	0.968	0.966	0.579	0.552	48
pl	0.542	0.542	0.642	0.642	0.579	0.579	0.500	0.500	0.934	0.934	0.529	0.529	48
it	0.590	0.609	0.675	0.679	0.581	0.600	0.497	0.520	0.954	0.914	0.574	0.587	48
fr	0.520	0.501	0.596	0.567	0.533	0.508	0.459	0.424	0.934	0.925	0.502	0.494	48
tr	0.457	0.460	0.525	0.533	0.496	0.490	0.450	0.433	0.907	0.912	0.435	0.441	48
hu	0.487	0.531	0.592	0.596	0.535	0.550	0.469	0.477	0.897	0.928	0.476	0.514	48
et	0.493	0.516	0.588	0.596	0.515	0.506	0.440	0.433	0.890	0.877	0.480	0.502	48
de	0.570	0.535	0.638	0.604	0.556	0.535	0.470	0.460	0.914	0.897	0.549	0.519	48
es	0.622	0.645	0.692	0.708	0.615	0.621	0.524	0.528	0.945	0.968	0.602	0.625	48
pt	0.628	0.637	0.721	0.717	0.627	0.621	0.520	0.517	0.957	0.957	0.614	0.622	48
AVG.	0.561	0.566	0.642	0.635	0.569	0.565	0.487	0.483	0.933	0.932	0.543	0.549	48
δ w.r.t. Baseline (%)	1.407	1.905	3.008	3.799	3.640	4.261	4.567	4.391	1.027	1.720	0.511	0.986	

Table 7. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from both Agrovoc and Organic.Lingua ontologies.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.691	0.691	0.754	0.754	0.671	0.671	0.570	0.570	0.973	0.973	0.659	0.659	48
el	0.554	0.574	0.654	0.625	0.560	0.548	0.475	0.484	0.936	0.958	0.539	0.560	48
lv	0.609	0.579	0.667	0.646	0.606	0.588	0.500	0.477	0.970	0.969	0.589	0.564	48
pl	0.557	0.557	0.675	0.675	0.606	0.606	0.502	0.502	0.943	0.943	0.537	0.537	48
it	0.617	0.620	0.708	0.696	0.608	0.617	0.501	0.525	0.952	0.942	0.601	0.601	48
fr	0.538	0.525	0.608	0.579	0.556	0.521	0.466	0.437	0.949	0.938	0.512	0.515	48
tr	0.481	0.481	0.550	0.550	0.515	0.508	0.457	0.441	0.917	0.926	0.465	0.459	48
hu	0.506	0.562	0.600	0.629	0.565	0.573	0.480	0.492	0.931	0.952	0.495	0.535	48
et	0.512	0.532	0.613	0.613	0.527	0.525	0.448	0.453	0.928	0.924	0.485	0.519	48
de	0.580	0.547	0.642	0.617	0.571	0.554	0.476	0.471	0.926	0.937	0.564	0.541	48
es	0.627	0.649	0.700	0.717	0.629	0.633	0.530	0.531	0.955	0.969	0.604	0.626	48
pt	0.642	0.648	0.721	0.717	0.640	0.633	0.521	0.516	0.962	0.966	0.634	0.634	48
AVG.	0.576	0.580	0.658	0.651	0.588	0.581	0.494	0.491	0.945	0.950	0.557	0.562	48
δ w.r.t. Baseline (%)	4.172	4.582	5.514	6.410	7.021	7.337	5.958	6.190	2.298	3.642	3.124	3.553	

too, due to the same reason explained in the previous section. Therefore, queries are expanded by exploiting (i) only the Organic.Lingua ontology (Table 8); or (ii) both ontologies (Table 9). In both cases, the query result considers the field containing manual annotations (that refer only to Organic.Lingua concepts) with a double weight.

By considering the improvements obtained with the usage of the manual annotations, we performed experiments by boosting the fields containing the manual annotations in order to verify if further improvements are obtained. Unfortunately, it seems that boosting these fields does not lead to any improvement. Indeed, except for the Prec@5 and the Prec@10 values, we registered a general decrease of the improvements with respect to the results obtained by running the previous configuration. Moreover, we may observe the only negative value with respect to the baseline of the entire evaluation, that

Table 8. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from Organic.Lingua ontology. The fields containing manual annotations have been weighted double.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.670	0.670	0.742	0.742	0.648	0.648	0.554	0.554	0.972	0.972	0.635	0.635	48
el	0.542	0.562	0.642	0.608	0.552	0.544	0.457	0.470	0.928	0.935	0.522	0.545	48
lv	0.597	0.575	0.654	0.646	0.592	0.583	0.490	0.466	0.968	0.966	0.573	0.545	48
pl	0.536	0.536	0.633	0.633	0.579	0.579	0.493	0.493	0.934	0.934	0.524	0.524	48
it	0.584	0.601	0.675	0.675	0.583	0.600	0.490	0.513	0.954	0.914	0.562	0.575	48
fr	0.519	0.499	0.592	0.567	0.533	0.506	0.458	0.424	0.934	0.925	0.503	0.496	48
tr	0.458	0.463	0.533	0.542	0.492	0.492	0.449	0.434	0.907	0.912	0.443	0.448	48
hu	0.480	0.525	0.596	0.588	0.529	0.552	0.458	0.469	0.897	0.928	0.467	0.509	48
et	0.488	0.511	0.583	0.596	0.517	0.510	0.434	0.424	0.890	0.877	0.473	0.492	48
de	0.570	0.536	0.633	0.604	0.558	0.538	0.469	0.459	0.914	0.897	0.547	0.517	48
es	0.618	0.639	0.688	0.708	0.615	0.617	0.516	0.519	0.945	0.968	0.595	0.613	48
pt	0.622	0.631	0.721	0.717	0.631	0.623	0.513	0.508	0.957	0.957	0.606	0.611	48
AVG.	0.557	0.562	0.641	0.636	0.569	0.566	0.482	0.478	0.933	0.932	0.538	0.543	48
δ w.r.t. Baseline (%)	0.693	1.306	2.820	3.783	3.580	4.460	3.361	3.186	1.010	1.719	-0.524	-0.138	

is the Avg. R-Precision. However, in spite of lower performance values of the sole application of the Organic.Lingua ontology, the improvements obtained by combining the two ontologies still remain in line with the ones obtained without boosting the manual annotations fields.

5.4 Use of Ontology Concepts Only

Here, queries are performed only on document fields containing ontological annotations. For this experiment, we have explored all three variants; queries have been performed (i) on fields containing Agrovoc annotations (Table 10); (ii) on fields containing Organic.Lingua annotations (both automatic and manual annotations) (Table 11); and (iii) on fields containing Agrovoc or Organic.Lingua annotations (Table 12). In this case, only queries containing at least one term matching the Domain Ontology have been performed.

From an initial glance, we may notice that in the results obtained with this configuration not all queries were able to be performed because, for some of them, no matches have been found in the respective ontologies. For instance, if we consider the Estonian language (that, by the way, is available only in the Organic.Lingua ontology) only for 6 queries there were found matches between query terms and ontology concepts. Moreover, not all languages are available for all ontologies. Indeed, Agrovoc ontology covers 9 out 12 languages; while, Organic.Lingua ontology covers 10 out of 12 languages. These two aspects confirm what we have already described previously in the paper where we stated that one of the main problems in using semantics and multilinguality for indexing and retrieving purposes is the non-complete coverage of the language terms and, sometimes, the unavailability of all languages in the semantic resources.

Besides this, it is anyway interesting to observe the results obtained by performing queries using only the ontological concepts that match query terms. As we expected, the results obtained by using the sole Organic.Lingua ontology outperforms both other

Table 9. Results obtained by performing queries using the machine translation service enriched with the matched URIs coming from both Agrovoc and Organic.Lingua ontologies. The fields containing manual annotations have been weighted double.

Lang	MAP		Prec@5		Prec@10		Prec@20		Avg. Recall		Avg. R-Prec.		Query Num.
	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	BING	GOOG	
en	0.688	0.688	0.750	0.750	0.671	0.671	0.570	0.570	0.973	0.973	0.659	0.659	48
el	0.553	0.573	0.646	0.621	0.563	0.548	0.475	0.484	0.936	0.958	0.538	0.558	48
lv	0.606	0.577	0.663	0.642	0.602	0.588	0.499	0.478	0.970	0.969	0.587	0.564	48
pl	0.552	0.552	0.667	0.667	0.602	0.602	0.503	0.503	0.943	0.943	0.535	0.535	48
it	0.609	0.616	0.700	0.692	0.600	0.613	0.501	0.524	0.952	0.942	0.593	0.598	48
fr	0.522	0.511	0.596	0.579	0.552	0.523	0.463	0.440	0.949	0.938	0.498	0.502	48
tr	0.462	0.463	0.529	0.533	0.506	0.500	0.451	0.434	0.917	0.926	0.444	0.439	48
hu	0.496	0.548	0.596	0.625	0.560	0.571	0.479	0.492	0.931	0.952	0.485	0.521	48
et	0.510	0.529	0.604	0.604	0.533	0.525	0.449	0.454	0.928	0.924	0.483	0.517	48
de	0.578	0.546	0.638	0.613	0.569	0.552	0.476	0.471	0.926	0.937	0.564	0.541	48
es	0.625	0.648	0.696	0.713	0.629	0.633	0.530	0.531	0.955	0.969	0.604	0.626	48
pt	0.639	0.646	0.713	0.713	0.635	0.629	0.524	0.520	0.962	0.966	0.631	0.631	48
AVG.	0.570	0.575	0.650	0.646	0.585	0.580	0.493	0.492	0.945	0.950	0.552	0.558	48
δ w.r.t. Baseline (%)	3.043	3.543	4.237	5.498	6.507	6.967	5.847	6.210	2.291	3.647	2.113	2.638	

Table 10. Results obtained by performing queries using only the terms matching concepts defined in the Agrovoc ontology

Lang	MAP	Prec@5	Prec@10	Prec@20	Avg. Recall	Avg. R-Prec.	Query Numbers
en	0.139	0.156	0.180	0.177	0.681	0.146	45
pl	0.115	0.171	0.183	0.159	0.509	0.139	35
it	0.140	0.195	0.195	0.191	0.546	0.168	38
fr	0.110	0.145	0.145	0.155	0.506	0.131	40
tr	0.122	0.171	0.181	0.164	0.531	0.145	42
hu	0.160	0.216	0.214	0.196	0.578	0.177	37
de	0.150	0.231	0.213	0.209	0.461	0.181	32
es	0.145	0.214	0.202	0.186	0.612	0.166	42
pt	0.153	0.200	0.205	0.178	0.544	0.174	43
AVG.	0.137	0.189	0.191	0.179	0.552	0.159	

configurations. Indeed, while the Agrovoc ontology is used only for the automatic annotation of documents, the Organic.Lingua one is exploited also for performing manual annotations. This further enrichment of the documents representation permits to increase the effectiveness of the CLIR system.

However, the combined use of the two ontologies is destructive with respect to the use of the sole Organic.Lingua one. We may notice that the number of queries matched by the two ontologies is different, and, from a more in depth analysis, we observed that some of the queries contains only partial matches with the Agrovoc concepts; while, by considering the Organic.Lingua one, no matches are found. This fact, even if it permits to handle more queries, it introduces in the evaluation results that reduce the overall effectiveness of the system.

5.5 General Remarks

Summarizing what we observed in our experiments, we may state that the use of domain-specific multilingual resources for enriching basic CLIR systems leads to

Table 11. Results obtained by performing queries using only the terms matching concepts defined in the Organic.Lingua ontology

Lang	MAP	Prec@5	Prec@10	Prec@20	Avg. Recall	Avg. R-Prec.	Query Numbers
en	0.267	0.381	0.343	0.338	0.683	0.283	21
el	0.288	0.381	0.367	0.345	0.640	0.302	21
lv	0.079	0.100	0.100	0.100	0.513	0.095	14
it	0.242	0.300	0.325	0.317	0.663	0.267	12
fr	0.218	0.300	0.236	0.214	0.598	0.240	14
tr	0.261	0.427	0.368	0.336	0.552	0.285	22
hu	0.350	0.471	0.421	0.454	0.677	0.373	14
et	0.243	0.333	0.233	0.308	0.741	0.263	6
de	0.333	0.491	0.436	0.473	0.679	0.381	11
es	0.324	0.427	0.382	0.348	0.654	0.341	22
AVG.	0.260	0.359	0.319	0.322	0.635	0.283	

Table 12. Results obtained by performing queries using only the terms matching concepts defined in the Agrovoc or in the Organic.Lingua ontologies

Lang	MAP	Prec@5	Prec@10	Prec@20	Avg. Recall	Avg. R-Prec.	Query Numbers
en	0.176	0.236	0.242	0.232	0.700	0.179	45
el	0.288	0.381	0.367	0.345	0.640	0.302	21
lv	0.079	0.100	0.100	0.100	0.513	0.095	14
pl	0.115	0.171	0.183	0.159	0.509	0.139	35
it	0.143	0.200	0.197	0.194	0.561	0.173	39
fr	0.126	0.181	0.171	0.167	0.528	0.143	41
tr	0.166	0.262	0.248	0.216	0.561	0.179	42
hu	0.193	0.268	0.242	0.241	0.610	0.201	38
et	0.243	0.333	0.233	0.308	0.741	0.263	6
de	0.202	0.303	0.279	0.280	0.545	0.234	33
es	0.215	0.307	0.274	0.249	0.672	0.223	43
pt	0.153	0.200	0.205	0.178	0.544	0.174	43
AVG.	0.173	0.247	0.226	0.221	0.586	0.192	

effective results. Indeed, in all experiments performed on our document collection, the usage (sole or combined) of the two described ontologies outperformed our baseline.

It is important to highlight also that the used baselines represent two of the most important and effective translation systems currently available. With respect to what we discussed previously in the paper, these baselines systems have been built by using dictionaries that almost completely cover each language. By comparing the proposed approach with them, it presents at least two important benefits: (i) the problem of building an effective machine translation system is demanded to external services and (ii) different ontologies, based on the domain/s that the system has to cover, may be plugged in order to improve its effectiveness.

By analyzing the results, we may observe that the major improvements are visible for the Prec@5 and Prec@10 values. This result demonstrates the feasibility of the approach that is able to improve the rank of the traditional first page results of information retrieval systems.

Concerning recall values, we may notice that the baselines already obtain significant recall values and the improvements obtained by adopting the ontologies are quite limited. This is an interesting point because it demonstrates that, even if we use a domain-specific scenario, the adopted baselines performed well during translation operations

because relevant documents are not lost during the retrieval phase. Therefore, we have a further evidence that the use of the ontologies for supporting general purpose machine translation services boosted the quality of the produced ranks.

However, we have also seen that the use of manual annotation significantly improves the results: around 7% versus around 4% for the automatic annotations. Moreover, if we observe the results obtained by performing queries containing only the ontology concepts, the use of the Organic.Lingua ontology (for which manual annotations are provided) leaded to significant better results (Table 11). Obviously, on the one hand it is almost well-known that the use of manual annotations improves the effectiveness of retrieval systems, but on the other hand, it requests a significant effort for keeping the system updated and, in complex real-world applications where thousands or million of documents are managed, it is not feasible.

6 Future Work and Concluding Remarks

In this work, we have presented a CLIR system based on the combination of the usage of domain-specific multilingual ontologies (i) for expanding queries and (ii) for enriching document representation with the index in a multilingual environment. The goal of the presented study was the investigation on the effectiveness of integrating semantic domain-specific resources, like ontologies, into a CLIR context. The implemented approach has been applied to a document collection built in the context of the Organic.Lingua EU-funded project where documents are domain-specific and where they have been annotated with concepts coming from domain-specific ontologies. The results have shown that the use of domain-specific resources for enriching the document representation and for performing a semantic expansion of queries is a suitable approach for improving the effectiveness of CLIR systems.

References

1. Salton, G.: Automatic processing of foreign language documents. In: COLING (1969)
2. Nie, J.Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
3. Ballesteros, L., Croft, W.B.: Resolving ambiguity for cross-language retrieval. In: SIGIR, pp. 64–71. ACM (1998)
4. Aljlayl, M., Frieder, O.: Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In: CIKM, pp. 295–302. ACM (2001)
5. Liu, Y., Jin, R., Chai, J.Y.: A maximum coherence model for dictionary-based cross-language information retrieval. In: Baeza-Yates, R.A., Ziviani, N., Marchionini, G., Moffat, A., Tait, J. (eds.) SIGIR, pp. 536–543. ACM (2005)
6. Gao, J., Nie, J.Y.: A study of statistical models for query translation: finding a good unit of translation. In: Efthimiadis, E.N., Dumais, S.T., Hawking, D., Järvelin, K. (eds.) SIGIR, pp. 194–201. ACM (2006)
7. Fung, P., Lo, Y.Y.: An ir approach for translating new words from nonparallel, comparable texts. In: Boitet, C., Whitelock, P. (eds.) COLING-ACL, pp. 414–420. Morgan Kaufmann Publishers/ACL (1998)

8. Pirkola, A., Toivonen, J., Keskustalo, H., Järvelin, K.: Fite-trt: a high quality translation technique for oov words. In: Haddad, H. (ed.) SAC, pp. 1043–1049. ACM (2006)
9. Mandl, T., Womser-Hacker, C.: How do named entities contribute to retrieval effectiveness? In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 833–842. Springer, Heidelberg (2005)
10. Munteanu, D.S., Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora. In: Calzolari, N., Cardie, C., Isabelle, P. (eds.) ACL. The Association for Computer Linguistics (2006)
11. Al-Onaizan, Y., Knight, K.: Translating named entities using monolingual and bilingual resources. In: ACL, pp. 400–408. ACL (2002)
12. Jaleel, N.A., Larkey, L.S.: Statistical transliteration for english-arabic cross language information retrieval. In: CIKM, pp. 139–146. ACM (2003)
13. Li, H., Sim, K.C., Kuo, J.S., Dong, M.: Semantic transliteration of personal names. In: Carroll, J.A., van den Bosch, A., Zaenen, A. (eds.) ACL. The Association for Computational Linguistics (2007)
14. Kimura, F., Maeda, A., Hatano, K., Miyazaki, J., Uemura, S.: Cross-language information retrieval by domain restriction using web directory structure. In: HICSS, p. 135. IEEE Computer Society (2008)
15. Lu, W.H., Lin, R.S., Chan, Y.C., Chen, K.H.: Using web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. Decision Support Systems 45(3), 585–595 (2008)
16. Sacaleanu, B., Buitelaar, P., Volk, M.: A cross language document retrieval system based on semantic annotation. In: EACL, pp. 231–234 (2003)
17. Sorg, P., Cimiano, P.: Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data Knowl. Eng. 74, 26–45 (2012)
18. Aggarwal, N.: Cross lingual semantic search by improving semantic similarity and relatedness measures. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 375–382. Springer, Heidelberg (2012)
19. Braschler, M.: Combination approaches for multilingual text retrieval. Inf. Retr. 7(1-2), 183–204 (2004)
20. Dragoni, M., da Costa Pereira, C., Tettamanzi, A.: A conceptual representation of documents and queries for information retrieval systems by using light ontologies. Expert Syst. Appl. 39(12), 10376–10388 (2012)
21. Braschler, M., Peters, C.: Clef 2002 methodology and metrics. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 512–525. Springer, Heidelberg (2003)
22. Agosti, M., Di Nunzio, G.M., Ferro, N.: Scientific data of an evaluation campaign: Do we properly deal with them? In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 11–20. Springer, Heidelberg (2007)