Facilitating cross-language retrieval and machine translation by multilingual domain ontologies

Petr Knoth[*], **Trevor Collins**[*], **Elsa Sklavounou**[†], **Zdenek Zdrahal**[*]

[*] KMI, The Open University
Milton Keynes, United Kingdom
{p.knoth, t.d.collins, z.zdrahal}@open.ac.uk
[†] SYSTRAN
Paris, France
sklavounou@systran.fr

Abstract

This paper presents a method for facilitating cross-language retrieval and machine translation in domain specific collections. The method is based on a semi-automatic adaption of a multilingual domain ontology and it is particularly suitable for the eLearning domain. The presented approach has been integrated into a real-world system supporting cross-language retrieval and machine translation of large amounts of learning resources in nine European languages. The system was built in the context of a European Commission Supported project Eurogene and it is now being used as a European reference portal for teaching human genetics.

1. Introduction

A significant amount of research has been carried out in the NLP and Semantic Web technology fields in the last years. A few activities and projects, such as LT4eL (Lemnitzer et al., 2007) or LTfLL (LTfLL, 2008), have been launched with the objective to integrate these technologies with eLearning systems. One of the vital sub-objectives of these projects is to allow seamless access and retrieval of *multilingual* learning materials. In this paper we report on the activities undertaken in the context of *Eurogene (The First Pan-European Learning Service in the Field of Genetics)* project related to the problem of accessing and sharing multilingual learning resources.

More specifically, the article builds on the idea that eLearning systems should not only allow the cross-language retrieval of learning resources, but should be extended with machine translation capabilities to provide a better user experience. The proposed approach synchronizes the adaption of cross-language retrieval and machine translation in such a way that the performance of both systems improves. Although the presented method has been integrated into an eLearning system in the human genetics field, it is applicable in a broader context.

Many of the important players in the information retrieval field (including Google and Yahoo!) offer cross-language information retrieval (CLIR), some of them also provide machine translation (MT). While the performance of these systems is usually sufficient for general queries, CLIR and MT are often inaccurate for domain-specific queries. Large repositories storing domain specific content, such as PubMed which stores vast amounts of scholarly articles, have successfully adopted large thesauri/ontologies of domain terminology to improve the performance of their retrieval system (Lu et al., 2009). While there are efforts targeting cross-language retrieval in eLearning (Lemnitzer et al., 2007; Eichmann et al., 1998; Lu et al., 2008), the combination of the domain-specific retrieval and machine translation is rarely available.

Because of the low frequency of polysemy in domain specific collections, domain-specific MT systems are capable of achieving high performance. However, one of the main obstacles remain in the acquisition of terminology. At the same time, the domain terminology is usually an essential artefact used for query composition. Our method is motivated by this problem and tries to approach it by using a single terminological access point embodied by the multilingual domain ontology for both CLIR and MT. This allows to combine the strengths of ontology-based retrieval and domain-specific machine translation. In Section 2, approaches to domain CLIR with relation to MT are introduced. The theoretical foundation of the method for facilitating domain CLIR and MT is explained in Section 3. The application of the approach in the Eurogene system is then presented in Section 4 and the performance is discussed in Section 5. Finally, the contribution of the paper for the eLearning domain is summarized in Section 6.

2. Approaches to domain CLIR

There are two typical approaches to CLIR:

1. MT approach - The user's query is translated from the source language to the target language and submitted to the search system. This approach can be further divided into two cases:
   (a) MT of the query is performed and the query is submitted in all languages of interest.
   (b) A multilingual ontology is developed and used to map the submitted query to different languages.
2. Statistical approaches - The system is trained on a collection of texts (usually parallel). The user's query is then mapped to a language independent document vector using approaches, such as Latent Semantic Indexing (LSI) (Dumais, 1997).

Approach 1(a) requires the search system to be well-adapted for the translation of the terminology of the tar-

get domain. Depending on the MT system in hand, domain adaption is rule or statistically based. Rule-based approaches allow specifying rules expressing that a given term t_{L_1} in language L_1 corresponds to term t_{L2} in L_2. Statistical approaches to machine translation support automatic learning of such pairs from parallel corpora. Approach 1(b) is motivated by the fact that monolingual domain ontologies can be employed to improve the performance of the retrieval system by query expansion leveraging the ability of ontologies to represent synonyms linked to a concept and the hierarchical structure of concepts. Monolingual ontologies can be extended to multilingual ontologies.

Approach 2 is influenced by the size of the available parallel corpora which is critical for the performance of the retrieval system. The approach is, in general, more suitable for bilingual cross-language retrieval as it is usually difficult to find experts to build a domain-specific training set that would contain parallel texts from each language of interest to a common interlingua.

3. Synergy of CLIR and MT

Our method is based on the assumption that when we start to build a domain-specific system for sharing language resources, the amount of parallel corpora available is often limited. Our methodology uses a multilingual domain ontology as we argue that ontologies are well-suited for domain CLIR and can also be used for the adaption of the machine translation system. We presume an IR system and a MT system to be available. More specifically, our approach requires a hybrid MT system combining rule-based and statistical-based MT.

The method consists of two phases, which will be discussed in this section in detail: the *initialization phase* and the *bootstrapping phase*. The initialization phase takes as the input a collection of domain texts or an existing monolingual domain ontology and produces as an output a lightweight multilingual ontology of the target domain. While this step is performed just once, the bootstrapping phase is repeated as many times as necessary. The bootstrapping phase takes as the input the multilingual ontology produced in the initialization phase and adapts the MT system by extracting domain specific translation rules from the ontology. As the amount of learning resources stored in the system systematically grows, a statistical module of the MT system can be applied at any time to extract bilingual pairs of domain terms from the available collection of learning resources. These pairs are then used to semi-automatically enrich the multilingual ontology, thus improve the performance of the CLIR and later also the MT system.

The **initialization phase** can be further divided into:

1. Development of a *seed* monolingual ontology.
2. Extension of the ontology to multiple languages.

The **first step** of our approach requires building a small monolingual domain ontology of concepts. For our purposes, we will define the monolingual ontology as a quadruple $O = \langle C, T, E, f \rangle$, where C is a set of concepts (cognitive units of meaning - abstract ideas or mental symbols), T is a set of terms (textual representations of concepts), E is a set of oriented relations (*is-a* relations), such that $\langle C, E \rangle$ is a directed acyclic graph, and $f : T \rightarrow C$ is a surjective function from terms to concepts. Note that this implies that polysemy cannot be represented in our ontology. This is for our purposes intentional as we comprehend a domain as an area or part of an area in which the terminology is unambiguous.[1]. Today, lightweight ontologies can be built by reusing existing ontologies or by applying NLP methods for term extraction and ontology learning (Cimiano and Völker, 2005).

In the **second step**, the initial domain ontology is translated using MT and is validated by domain experts. The accuracy of MT is at this moment usually low as the system has not yet been sufficiently trained for the target domain. The resulting multilingual ontology is a 6-tuple $O = \langle C, T, E, f, L, lang \rangle$, where L is the set of languages and $lang : T \rightarrow L$ is a mapping from terms to languages. After the validation, the multilingual ontology is integrated with the retrieval system and the available collection of language resources is indexed in terms of the ontology. A set of terms $\{t | lang(t) = \text{language of the resource}\}$ is used for indexing.

The **bootstrapping phase** can be iterated as many times as necessary. The mutual updating procedure is shown in Figure 1. This phase can be further divided into:

1. Adaption of the MT dictionaries
2. Adaption of the multilingual ontology

In the **first step** of the bootstrapping phase, the MT system is adapted to the domain using bilingual substitution rules of form $t_{L_1} \rightarrow t_{L_2}$ extracted from the multilingual ontology and satisfying the condition $f(t_{L_1}) = f(t_{L_2})$, where $t_{L_1} \in T_{L_1}, t_{L_2} \in T_{L_2}$ and T_{L_n} is defined as $T_{L_n} = \{t | lang(t) = L_n\}$. For MT systems that translate using an interlingua, the term on the left hand side of a rule is a term in the language of the interlingua and the term on the right hand side is a term in any other supported language. For bilingual MT systems all combinations of terms are exploited and used for the generation of the translation rules. Supplying MT with rules extracted from the ontology can be also useful when a domain is accessed from a general-purpose search engine. IR systems can be equipped with a classification component that can: calculate the most probable domain of a document, select the most suitable domain ontology available, and extract the rules for adaption of the MT system.

For the **second step** of the bootstrapping phase, let us assume that the content stored in our system grows over time. Each time a new learning resource is submitted, it is indexed and put into the document collection. The submitted learning resource may be a translation of an already existing resource stored in the collection. Such parallel texts can be automatically recognized (Resnik and Smith, 2003) and used by the machine translation system for training.[2]

[1] Note that this assumption is not always true.

[2] Most of the statistical MT systems require parallel corpora

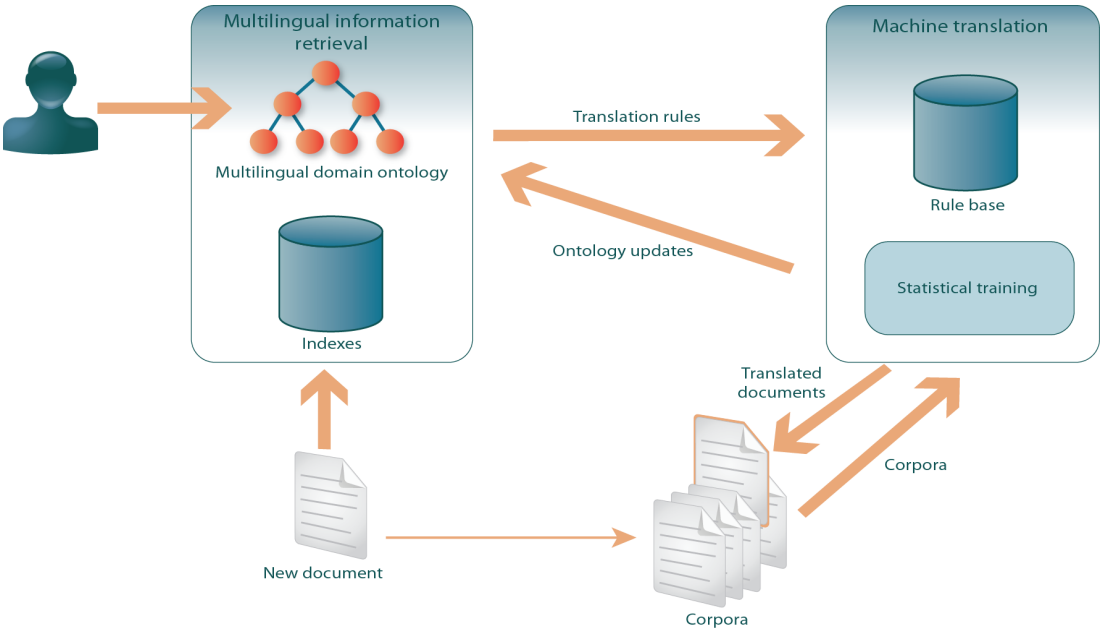

Figure 1: Collaboration of CLIR and MT. Translation rules are extracted from the multilingual ontology and are used to adapt the MT system. New terminology discovered in the statistical training phase is sent to the CLIR system which adapts the multilingual ontology. The updates are validated by a domain expert.

The output of the statistical training is a set of quadruples of the form $(t_{L_1}, t_{L_2}, conf, lang_q)$, where $conf$ is the confidence measure of translating term t_{L_1} to t_{L_2} estimated from text and $lang_q : T \rightarrow L$ is a mapping from terms to languages. The statistical model of the MT system is updated and the quadruples are sent to the CLIR system which uses the following algorithm to update the ontology:

Algorithm: Update ontology

Input: Multilingual ontology $O = \langle C, T, E, f, L, lang \rangle$, a set Q of quadruples of form $(t_{L_1}, t_{L_2}, conf, lang_q)$.

Output: An updated ontology $O' = \langle C, T', E, f', L, lang' \rangle$.

1. $T' := T, f' := f, lang' := lang, \tau :=$ arbitrary value from $[0, 1]$
2. for each $(t_{L_1}, t_{L_2}, conf, lang_q) \in Q$ do
3. if $lang_q(t_{L_1}) \in L \wedge lang_q(t_{L_2}) \in L \wedge conf \geq \tau$ then
4. if $t_{L_1} \in T_{L_1} \wedge t_{L_2} \notin T_{L_2}$ then
5. $T' := T' \cup t_{L_2}$
6. $f' := f' \cup (t_{L_2}, f(t_{L_1}))$
7. $lang' := lang' \cup (t_{L_2}, lang_q(t_{L_2}))$
8. end if
9. if $t_{L_2} \in T_{L_2} \wedge t_{L_1} \notin T_{L_1}$ then
10. $T' := T' \cup t_{L_1}$
11. $f' := f' \cup (t_{L_1}, f(t_{L_2}))$
12. $lang' := lang' \cup (t_{L_1}, lang_q(t_{L_1}))$
13. end if
14. end if
15. end for
16. return $O' = \langle C, T', E, f', L, lang' \rangle$

The algorithm requires one pass through the set of quadruples Q (line 2). During initialization a sufficiently high value of parameter τ is set (line 1). Each quadruple is first tested for the compatibility with the ontological language set and for its confidence (line 3). Later, it is checked whether the terms suggested by MT can be mapped to the ontology (lines 4 and 9). The ontology is then updated using the components of the quadruple (lines 5-7 and 10-12). Finally, the algorithm assembles the new ontology (line 16). When the ontology is updated, domain terminology administrators are made aware of the updates by the system and, if necessary, modifications can be performed (for example, new concepts should be added or better translation than the one proposed exists). Performed validation causes new pairs of rules $t_{L_1} \rightarrow t_{L_2}$ to be extracted from the validated part of the ontology and to be submitted back to the rule base of the MT system. As the amount of content grows, the system bootstraps and the performance of both MT and CLIR is improved.

4. Application in human genetics

In this section, we describe an application of the method of Section 2 in the context of the Eurogene project, which provides an eLearning system for sharing learning resources in human genetics.[3] The learning resources are submitted to the system typically in the form of slides, books and research articles represented in a variety of formats including Portable Document Format, Word, Power Point and many others. The Eurogene system also supports multimedia resources, such as images and videos in a number

for training, however there have been research studies that investigated learning of multilingual terminology from non-parallel texts, such as in (Fung and Mckeown, 1997).

[3]The system can be freely accessed at http://eurogene.open.ac.uk/

of formats. Resources can be handled in nine European languages[4], which are English, German, French, Spanish, Italian, Greek, Dutch, Czech and Lithuanian. More than 30 universities and other institutions located mainly across Europe, but also in non-European countries are actively contributing to this collection.

In Eurogene, the initial genetic ontology was developed by merging six monolingual ontologies[5] that contained a descriptive, but not too extensive, terminology of the domain. This ontology was translated into the above nine European languages (English is used as an interlingua, i.e. it is used to label the names of concepts) by domain experts and an upper-level ontology has been inferred using Unified Medical Language System (UMLS). A more comprehensive description of the ontology building process can be found in (Zdrahal et al., 2009).

The upper-level ontology helps to organize concepts from a relatively flat structure into a concept hierarchy, which is represented in the Simple Knowledge Organization System (SKOS) format which satisfies our definition of the ontology from the previous section. Figure 2 shows how a genetic concept *linkage analysis* is represented in our ontology.

The multilingual ontology was then integrated with the CLIR system. Since then, available content is being annotated. Textual resources are annotated automatically, multimedia resources are annotated manually, but the annotation procedure is guided by the ontology.

In the first part of the bootstrapping phase, rules were extracted from the multilingual ontology to adapt the MT system as described in the previous section. This typically helps to improve the performance of MT. For example, before the adaption, our system wrongly translated the English collocation *linkage analysis* to French as *analyse de triglerie*, whereas since the rule *Linkage analysis* $\rightarrow$ *Analyse de liasion* was extracted from the part of the ontology in Figure 2 and it was put into the MT rule base, the system has correctly translated the term as *Analyse de liasion*.

The CLIR system is powered by Lucene extended with a dedicated query parser that allows the user to combine terminological and full-text queries. Queries can be expressed in any of the available languages, and the results can be filtered by a subset of the available languages. Queries are mapped to a language independent representation using the ontology. The CLIR system can also be used during query composition to visualize the concept hierarchy and to interactively control query expansion for broader and/or narrower terms (Figure 3), thus utilizing the benefits of ontology-based retrieval.

A hybrid system developed by SYSTRAN is used for MT tasks, i.e. for the MT of resources and also for the learning of relations from parallel texts (SYSTRAN, 2009). The

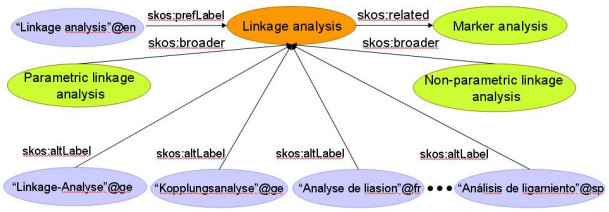

Figure 2: Representation of a concept *linkage analysis* in the multilingual ontology. The preferred label of this concept is the English version *Linkage analysis*. The concept has a two alternative representations in German (*Linkage-Analyse* and *Kopplungsanalyse*).[7] The representation in French is *Analyse de liasion* and in Spanish *Analisis de ligamiento*. The concept Linkage analysis is a broader concept for *Parametric linkage analysis* and *Non-parametric linkage analysis*, and it is related to a concept *Marker analysis*.

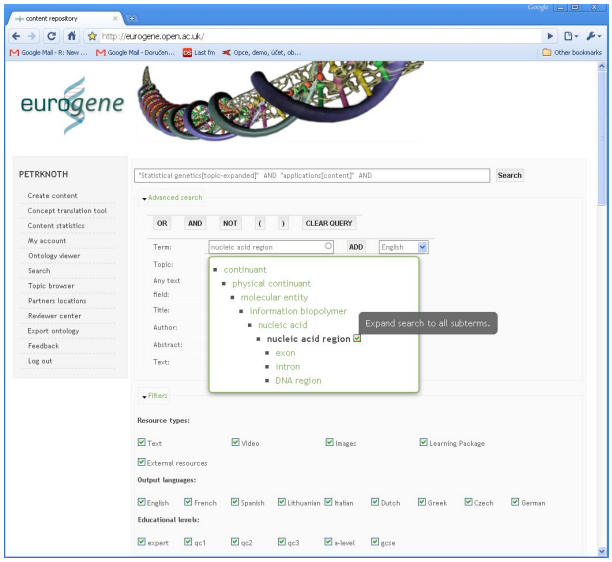

Figure 3: User interface of the Eurogene CLIR system. The CLIR system allows to control the expansion for broader/narrower terms.

CLIR and MT systems communicate using SOAP messages that allow the sending of extracted translation rules from CLIR to MT, and the sending of newly proposed translations from MT to CLIR. When newly proposed translations are received by CLIR, the ontology is updated using the algorithm in Section 2. Domain experts then perform terminology validation which is supported by the system and results in sending new translation rules to the MT rule base. This synchronization provides a mechanism for continuous semi-automatic adaption of both CLIR and MT systems.

5. Performance analysis

The performance of the proposed method and its impact on the resulting CLIR and MT systems can be influenced by a number of factors. These include mainly the suitability of the multilingual ontology for the target domain,

[4]While CLIR allows to pose queries and receive results in any of the mentioned languages, MT is limited to language pairs supported by the Systran system. Please also note that MT is not applied to images and videos.

[5]Published by the University of Washington in Seattle, National Institute of General Medical Sciences in Bethesda, Elsevier, Oracle ThinkQuest, University of Michigan and Centre for Genetics Education in Sydney

the amount of domain corpora available in the statistical phase, the performance of the multilingual keyword extraction system and the validity of the judgements performed by domain experts in the ontology refinement process. Given the number of possible error sources, it seems much more sensible to make sure that the method satisfies certain properties rather than performing a quantitative evaluation that would be biased by too many components.

One of the important properties that the proposed method in Section 3 should have is that the performance of both CLIR and MT should never decrease as a result of any bootstrapping iteration. Let us assume that the initial ontology has been validated by domain experts, so that it does not include any spurious translations. There are now two tasks which could have a negative impact on the performance of the CLIR or MT systems. These tasks correspond to 1) the update of the MT rule base and 2) the update of the multilingual ontology as described in Section 3.

If we assume that our domain is sufficiently small, so that no domain specific term appearing in the multilingual ontology is polysemous in our collection, then updating the dictionary of the MT system may either improve or not change the precision of the MT system. Since it is not possible to extract a spurious translation rule from the multilingual ontology, the resulting MT system cannot perform worse than before the update.

It is essential to expect that the statistical training phase described in Section 3 may produce quadruples describing translations that are in fact invalid and may thus introduce errors to the ontology. However, since all the updates must be validated by domain experts before they can be used by the CLIR system, it is possible to assume that no errors are introduced. This is in reality difficult as humans are in fact vulnerable to introducing errors. Thus, the quality of the ontology used by CLIR can deteriorate only under the condition that an error has been introduced by a domain expert.

To summarize, if all the above mentioned conditions are met, the method is guaranteed to improve or in the worst case not to worsen the performance of the CLIR and MT systems after each iteration.

6. Implications for eLearning

This paper showed that current eLearning applications supporting CLIR can also easily adopt MT and tailor it for their domain. In addition, the synergy of CLIR and MT may help to improve the performance of both. The main reason why the method is particularly useful in eLearning is that we should expect that the users of eLearning applications will very often use domain terminology as a part of their submitted queries, thus the added value will become more noticeable than in other contexts.

The paper brought the following contribution:

- Development of a new method for facilitating cross-language retrieval and machine translation by multilingual domain ontologies.
- Development of a real-world eLearning application enhanced by the use of the presented method.

7. Conclusion

Multilingual ontologies are particularly suitable for domains where terminology is used for query composition, such as in eLearning. They can be used as a synchronization component for domain adaption of CLIR and MT systems. In addition, the solution is easily readable and adjustable by humans and does not preclude the use of statistical approaches for terminology extraction when a large corpora is available. In the future, publishing of multilingual ontologies on the Web in a standard format may allow an application to decide which domain ontology to use for query expansion and for adaption of the MT system based on the context of the query. This may be helpful when a user accesses a specific domain from a general-purpose search engine.

8. References

Philipp Cimiano and Johanna Völker. 2005. Text2onto - a framework for ontology learning and data-driven change discovery.

Susan T. Dumais. 1997. Automatic cross-language retrieval using latent semantic indexing.

David Eichmann, Miguel E. Ruiz, and Padmini Srinivasan. 1998. Cross-language information retrieval with the umls metathesaurus. In *In: Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 72–80.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora.

Lothar Lemnitzer, Cristina Vertan, Alex Killing, Kiril Ivanov Simov, Diane Evans, Dan Cristea, and Paola Monachesi. 2007. Improving the search for learning objects with keywords and ontologies. In Erik Duval, Ralf Klamma, and Martin Wolpers, editors, *EC-TEL*, volume 4753 of *Lecture Notes in Computer Science*, pages 202–216. Springer.

LTfLL. 2008. Language technology for lifelong learning (ltfll).

Wen-Hsiang Lu, Ray S. Lin, Yi-Che Chan, and Kuan-Hsi Chen. 2008. Using web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. *Decis. Support Syst.*, 45(3):585–595.

Zhiyong Lu, Won Kim, and W. John Wilbur. 2009. Evaluation of query expansion using mesh in pubmed. *Inf. Retr.*, 12(1):69–80.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

SYSTRAN. 2009. Systran's machine translation technology url: http://www.systran.co.uk/systran/corporate-profile/translation-technology.

Zdenek Zdrahal, Petr Knoth, Trevor Collins, and Paul Mulholland. 2009. Reasoning across multilingual learning resources in human genetics. In *Proceedings of ICL 2009*.