Ontology-Based Multilingual Information Retrieval

Jacques Guyot[*] Saïd Radhouani[*,**] Gilles Falquet[*]

[*]Centre universitaire d'informatique
24, rue Général-Dufour, CH-1211 Genève 4, Switzerland

[**]Laboratoire CLIPS-IMAG, B.P. 53, 38041 Grenoble cedex 9, France

Jacques.Guyot@rolex.com, {Said.Radhouani, Gilles.Falquet}@cui.unige.ch

Abstract. For our first participation in the CLEF evaluation campaign, our aim is to explore a translation-free technique for multilingual information retrieval. This technique is based on an ontological representation of documents and queries. We use a multilingual ontology for documents/queries representation. For each language, we use the multilingual ontology to map a term to its corresponding concept. The same mapping is applied to each document and each query. Then, we use a classic vector space model for the indexing and the querying. The main advantages of our approach are: no merging phase is required, no dependency on automatic translators between all pairs of languages exists, and adding a new language only requires a new mapping dictionary to the multilingual ontology.

Key words: Multilingual Ontology, Conceptual Indexing, Multilingual Information Retrieval.

Introduction

The existing approaches use either translation of all documents into a common language, either automatic translation of the queries, or combination of both query and document translations [Chen at al. 2003]. In these cases, we need automatic translators between all pairs of languages. If we translate queries, after receiving a result list from each search engine, we need to use a merging procedure to provide a unique ranked result list. Moreover adding a new language (query or document) requires as much translators as existing languages.

In our approach, we tried to "dissolve" these problems by using of a multilingual ontology. Based on this ontology, we conducted different experiments involving multilingual test-collection. We retrieve documents written in Dutch, English, Finnish, French, German, Italian, Spanish and Swedish, independently of the query language. First, we tried to prove the feasibility of our approach using English when submitting queries. We also tried to prove that our system is independent of the query language. Thus, we have used Dutch, French, and Spanish when submitting queries,

In the next section, we describe our approach and present our official runs.

1. Ontology based Multilingual Information Retrieval

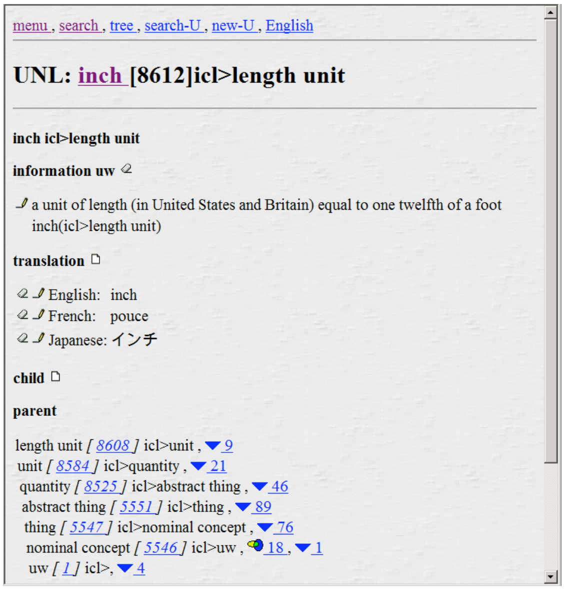
1.1 Multilingual ontology

A Multilingual ontology is defined by one ontology and a set of dictionary (one dictionary for each language). An ontology is a formal, explicit specification of a shared conceptualisation [Gruber 1993]. It contains a set of distinct and identified concepts ***C*** related by a set of relations ***R***. In our approach, we only need to use the set of concepts. Here, we present two examples of concepts extracted from our ontology:

- *8612 : a unit of length (in United States and Britain) equal to one twelfth of a foot*
- *28845: the thick short innermost digit of the forelimb.*

A dictionary $\boldsymbol{D_L}$ is an association of ontology concepts $\boldsymbol{C}$ with a terms set $\boldsymbol{T_L}$ pertaining to a language $\boldsymbol{L}$. We denote: $\boldsymbol{D_L : C \rightarrow T_L}$. Indeed, the concept c is labelled by a set of terms $t_1,t_2,..,t_n$ in the language L. We denote $D_L(c)=\{t_1,t_2, \ldots,t_n\}$. We also define the reciprocal relation S_L $\boldsymbol{: T_L \rightarrow C}$ by S_L $(t)=\{c\in C \mid t \in D_L(c)\}$. Actually, the term t indicates the concepts $c_1,c_2,\ldots, c_m$. We also denote $S_\mathrm{L}(t)=\{ c_1,c_2, \ldots,c_m\}$. Here, we present two examples of associations between terms and concepts:

- $D_{EN}(28845)= \{thumb\}$.
- $S_{FR}(pouce)= \{8612, 28845 \}$.



menu , search , tree , search-U , new-U , English

UNL: inch [8612]icl>length unit

inch icl>length unit

information uw

a unit of length (in United States and Britain) equal to one twelfth of a foot
inch(icl>length unit)

translation

English: inch
French: pouce
Japanese: インチ

child

parent

length unit [8608] icl>unit , ▼ 9
unit [8584] icl>quantity , ▼ 21
quantity [8525] icl>abstract thing , ▼ 46
abstract thing [5551] icl>thing , ▼ 89
thing [5547] icl>nominal concept , ▼ 76
nominal concept [5546] icl>uw , 18 , ▼ 1
uw [1] icl>, ▼ 4

Figure 1. Example of concept definition in UNL[UNL].

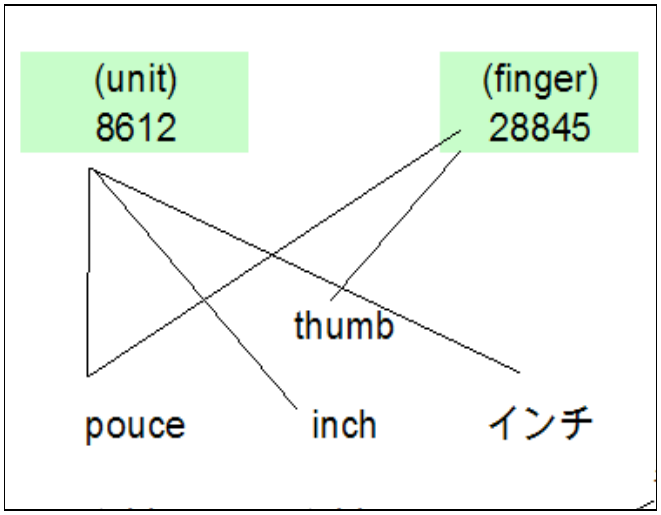


Figure 2. Example of term-concept association.

To bootstrap and build dictionaries, we have used Esperanto dictionaries found on the Web (principally from Ergane [ERG 2005]). We have also used automatic translations to complete some of them.

Language	Stemmer	#concepts (human check)	Add automatic translations
English	eng	5300	
French	fra	13000	
German	deu	20000	x
Dutch	dut	81000	x
Italian	ita	4500	x
Spanish	spa	19000	x
Swedish	sve	6500	x
Finnish	fin	2000	

Tableau 1. Description of the linguistic used resources.

1.2 Ontology based multilingual information retrieval

In our approach, for each document in the whole collection, we use the multilingual ontology to map each term to its corresponding concept. We apply the same process on the queries.

The document $d_L = < t_1, t_2,\ldots, t_n>$ is a sequence of terms from the set T_L of the language L. To carry out the term-concept mapping, we apply the function S_L on each term t_i of the document d_L: $S_L(t)=\{c \in C \mid t \in D_L(c)\}$. So we obtain the conceptual representation of the document that we denote: $CR(d_L)= < S(t_1), S(t_2), \ldots, S(t_n)>$. Finally, $CR(d_L)$ is a sequence of sets of concepts.

We did not introduce any treatment for the term ambiguity. In fact, if the term is ambiguous, we replace it by all its corresponding concepts.

Before the term mapping step, we use a "stop word list" for each language, and a dedicated stemming system. We have used *Snowball*, a small string processing language designed for creating stemming algorithms in Information Retrieval [Snow 2005].

We did not introduce any morpho-syntactic or processing (like n-grams) to break composite words in Dutch, German, or Finnish.

For indexing and querying, we use the vector space model [Salton et al. 83].

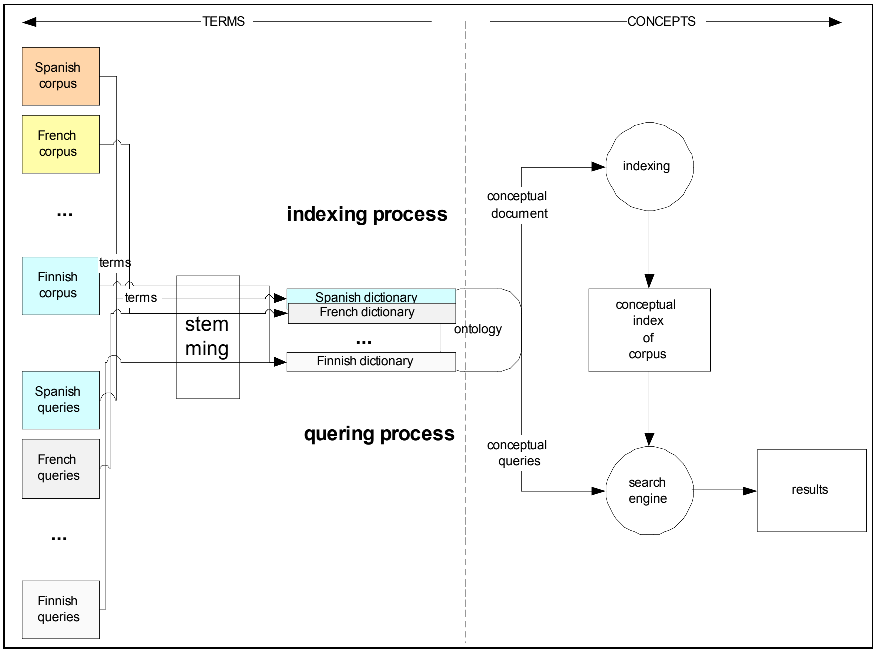


Figure 3. Indexing and querying process

1.3 Official runs description

In our approach, each query Q is composed by two fields: a topic field and a body field. We denote Q=<Topic,Body>. The content of each field depends on the runs. Each field is composed by a list of terms extracted from the original text query.

As the queries are precise, we use the topic field to query the whole collection. As a result, we obtain a set of documents containing the topic concepts. Then we use the body field to rank this set of documents.

Here we present an example of a query composed by a topic field (text between topic tags) and a body field (text between all the other tags).

<top>
<num> C182 </num>
<topic> Normandië Landing </topic>
<NL-title> 50e Herdenkingsdag van de Landing in Normandië </NL-title>
<NL-desc> Zoek verslagen over de dropping van veteranen boven Sainte-Mère-Église tijdens de viering van de 50e herdenkingsdag van de landing in Normandië. </NL-desc>
<NL-narr> Ongeveer veertig veteranen sprongen tijdens de viering van de 50e herdenkingsdag van de landing in Normandië met een parachute boven Sainte-Mère-Église, net zoals ze vijftig jaar eerder op D-day hadden gedaan. Alle informatie over het programma of over de gebeurtenis zelf worden als relevant beschouwd. </NL-narr>
</top>

Now we present our official runs. In the following three runs, we use English when submitting queries:

1. AUTOEN: the topic field is composed by the terms of the title of the original query (text between the title tags). The body field is composed by the text of the original query.
2. ADJUSTEN: the topic field is composed by the modified title by adding and/or removing terms. The adding terms are extracted from the original query text. The body field is composed by the text of the original query.
3. FEEDBCKEN: the topic field is composed by the modified title as in ADJUSTEN. The body field is composed by the original text query and the first relevant document (if it exists) in the first 30 documents found by the previous ADJUSTEN run.

Table 1 shows the result of each run. Of course it's difficult to have a good result for the AUTOEN run while the topic contains all the concepts corresponding to the query title terms. We succeeded in improving the result of 63.11% by using the adjusted topic. Finally, by using the relevance feedback, we improved the result of 24.74%. This improvement is due to the vector space model which gives better results when the documents/queries vectors are long.

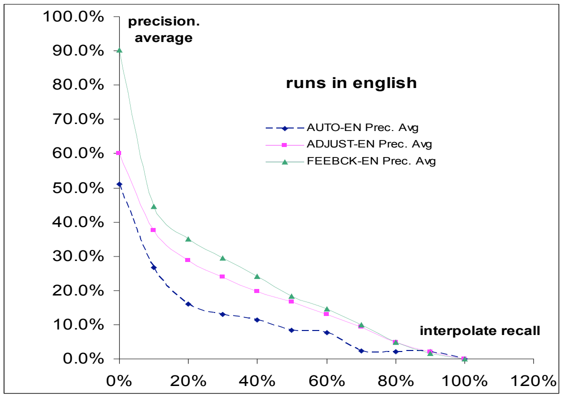


Figure 4. Comparison of the system result using three strategies

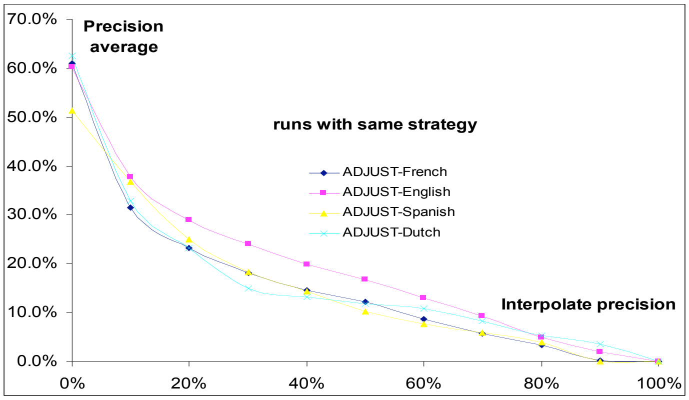


Figure 5. Comparison of the system result using four languages

In order to compare the system results using different languages when submitting queries, we carried out three more runs: ADJUSTDU, ADJUSTFR, and ADJUSTSP. For each run, we use respectively Dutch, French, and Spanish when submitting queries. In these runs, topic field is composed by modified title like in ADJUSTEN and body field is composed by the original query text.

For all the four runs, we obtain almost the same mean average precision: 13.90% for ADJUSTDU, 13.47% for ADJUSTFR, 13.80% for ADJUSTSP and 16.85 % for ADJSUTEN. Our system is not dependent of the query language. It gives nearly the same results when submitting queries in four different languages. It's difficult to explain difference because the coverage and the quality of ontological dictionaries are important.

Run name	Query language	Type	MAP
AUTO-EN	English	Automatic	10.33 %
ADJSUT-EN	English	Adjusted Topic	16.85 %
ADJUST-DU	Dutch	Adjusted Topic	13.90 %
ADJUST-FR	French	Adjusted Topic	13.47 %
ADJUST-SP	Spanish	Adjusted Topic	13.80 %
FEEDBCK-EN	English	Manual	**21.02** %

Table 2. Description and Mean Average Precision (MAP) of our official multilingual runs

Conclusion

In this CLEF evaluation campaign, we evaluated a multilingual ontology-based approach for multilingual information retrieval. We did not use any translation either for documents or for queries. We carried out a common document/query representation based on multilingual ontology. Then, we used the vector space model for indexing and querying. Compared with the existing approaches, our approach has several advantages. Indeed, there is no dependency on automatic translators between all pairs of languages. When we add a new language, we only add, in the ontology, a new mapping dictionary. Also, we do not need any merging technique to rank the list of retrieved documents.

In this preliminary work, we tried only to prove the feasibility of our approach. We tried also to prove that our system is independent of the query language. We still have some limits in our system because we did not introduce any morpho-syntactic processing to break composite words in Dutch, German, or Finnish. Moreover, our ontology is incomplete and dirty (we have imported many errors with automatic translation).

We have also used the same approach in the bi-text alignment field. We have used other language like Chinese, Arabic and Russian [Guyot 2005].

Acknowledgments

We would like to thank the CLEF-2005 organisers for their efforts. We would also like to thank Metaread for giving us the possibility to use the “idxvli” information retrieval system (A fast indexer for big corpora).

References

[Chen at al. 2003] Chen, A. and Gey, F. Combining query translation and document translation in cross-language retrieval. *In proceedings CLEF-2003*, pp. 39.48. Trondheim.

[ERG 2005] Ergane: http//download.travlang.com/, see also http://www.majstro.com/-

[Guyot 2005] GUYOT, J. yaaa: yet another alignment algorithm - Alignement ontologique bi-texte pour un corpus multilingue. Cahier du CUI 2005.

[Gruber 1993] Gruber, T. R. A translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5 : 199-220, 1993.

[UNL] UNL: Universal Networking Language. http://cui.unige.ch/isi/unl/ & http://www.undl.org/.

[Salton et al. 83] Salton, G. and Mcgill, M. J. Introduction to the modern Information Retrieval. McGraw-Hill (1983)

[Snow 2005] SnowBall: http://snowball.tartarus.org/.